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

UMA ARQUITETURA SUPORTADA POR BUSCA SEMÂNTICA PARA RECUPERAÇÃO DE FONTES DE INFORMAÇÃO EM REPOSITÓRIOS DE METADADOS

Veronica dos Santos

Orientadores
Asterio Kiyoshi Tanaka
Fernanda Araujo Baião

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

UMA ARQUITETURA SUPORTADA POR BUSCA SEMÂNTICA PARA RECUPERAÇÃO DE FONTES DE INFORMAÇÃO EM REPOSITÓRIOS DE METADADOS

Veronica dos Santos

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

Aprovada por:

Asterio Kiyoshi Tanaka, Ph.D. – UNIRIO

Fernanda Araujo Baião, D.Sc. – UNIRIO

Maria Luiza Machado Campos, Ph.D. – UFRJ

Sean Wolfgand Matsui Siqueira, D.Sc. – UNIRIO

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Santos, Veronica dos.
S237 Uma arquitetura suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados / Veronica dos Santos, 2011.
170f.

Orientador: Astério Kiyoshi Tanaka.
Coorientador: Fernanda Araujo Baião.
Dissertação (Mestrado em Informática) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2011.

1. Sistemas de informação. 2. Busca semântica. 3. Repositórios de metadados. 4. Sistemas de recuperação da informação. 6. Ontologias de domínio. I. Tanaka, Astério Kiyoshi. II. Baião, Fernanda Araujo. III. Universidade Federal do Estado do Rio de Janeiro (2003-). Centro de Ciências Exatas e Tecnologia Curso de Mestrado Informática. III. Título.

CDD – 004.2

DEDICATÓRIA

Ao meu esposo Marcelo,

À minha mãe Lurdinha,

Ao meu irmão Victor,

À minha sogra Nita,

Ao meu pai Miguelito,

Aos meus filhos emplumados e

À todos com quem compartilho a Minha Jornada.

"O Amor é a única coisa que cresce à medida que se reparte."

(Antoine de Saint-Exupéry)

AGRADECIMENTOS

Agradeço:

- Aos orientadores Asterio Kiyoshi Tanaka e Fernanda Araujo Baião pela dedicação ao longo de todo o desenvolvimento deste trabalho.
- Aos professores Leonardo Guerreiro Azevedo e Sean Wolfgand Matsui Siqueira pelas valiosas contribuições durante os seminários.
- À professora Maria Luiza Machado Campos por sua receptividade em participar da banca de avaliação.
- Aos demais professores do Programa de Pós-Graduação em Informática da UNIRIO pelo conhecimento adquirido.
- Aos funcionários de apoio do PPGI pela cordialidade. Um agradecimento em especial ao Jackson Durães, sempre prestativo.
- Aos colegas de turma. Lembrem-se: "*Se o corpo não agüenta, a moral sustenta.*"
- Aos que participaram das etapas de avaliação desta pesquisa. Sem pessoas não seria possível gerar este conhecimento.
- À equipe do SERPRO que me recebeu de braços abertos. Um agradecimento em especial ao Ednylton Franzosi, por seu entusiasmo.
- Aos amigos e iniciados Hesley Py e Lúcia Castro.
- Aos demais colegas do IBGE. Em especial ao meu gerente Paulo Bahia e ao meu coordenador José Luiz Thomaselli por terem viabilizado a minha licença capacitação.

"Cada um que passa em nossa vida, passa sozinho,
pois cada pessoa é única e nenhuma substitui outra.
Cada um que passa em nossa vida, passa sozinho,
mas não vai só nem nos deixa sós.
Leva um pouco de nós mesmos, deixa um pouco de si mesmo.
Há os que levam muito, mas há os que não levam nada.
Essa é a maior responsabilidade de nossa vida,
e a prova de que duas almas não se encontram ao acaso."

(Antoine de Saint-Exupéry)

"People can’t share knowledge if they don’t speak a common language"
(Davenport e Prusak 1997)

Santos, Veronica dos. **Uma arquitetura suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados.** UNIRIO, 2011. 170 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Integração de informações é um processo interativo e iterativo. Na maioria das organizações, diferentes fontes de informação coexistem e se sobrepõem em seu conteúdo, exigindo conhecimento de domínio para descobrir, compreender e integrar informações. A catalogação de fontes de informação em um repositório de metadados corporativo pode aumentar o seu potencial de reutilização e integração, mas metadados descritivos não são suficientes para permitir a descoberta consistente e inequívoca de quais fontes de informação contêm os dados necessários para integração. Esta pesquisa acadêmica propõe uma arquitetura lógica suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados. Essa arquitetura faz uso de ontologias de domínio para formalizar as relações semânticas entre os conceitos de um domínio particular, para recuperar o melhor conjunto de fontes de informação semanticamente similares e aumentar a semântica de metadados descritivos através de anotações semânticas. A avaliação daa arquitetura foi realizada através de dois métodos: um experimento e um estudo de caso. O experimento, usando um protótipo da arquitetura proposta, obteve resultados positivos no que diz respeito à precisão e cobertura em comparação com outras abordagens da literatura. O estudo de caso, realizado em uma aplicação do governo brasileiro, o Catálogo de Informações *DadosGov* COI-PR, demonstrou os benefícios e dificuldades da implementação desta arquitetura em um cenário do mundo real.

Palavras-chave: Integração Semântica de Informações; Busca Semântica; Ontologias de Domínio, Repositório de Metadados.

ABSTRACT

Information integration is an interactive and iterative process. In most organizations, different information sources coexist and their contents overlap, thus requiring domain knowledge for discovering, understanding and integrating information. Cataloging information sources in a corporate metadata repository can increase the potential for their reuse and integration; however, their descriptive metadata does not suffice for the consistent and unambiguous discovery of which information sources contain the required data to be integrated. This academic research proposes a logical architecture supported by semantic search for information sources retrieval from metadata repositories. This architecture makes use of domain ontologies to capture semantic relationships among concepts from a particular domain, to retrieve the best subset of semantically similar information sources and to augment descriptive metadata semantics through semantic annotations. The proposed architecture was evaluated according to two methods: an experiment and a case study. Our experiments, using a prototype of the proposed architecture, obtained positive results with regard to precision and recall compared to other approaches in the literature. The case study, carried out on a Brazilian government application, *DadosGov* COI-PR Information Catalog, demonstrated the benefits and difficulties on implementing this architecture in a real-world scenario.

Keywords: Semantic Information Integration; Semantic Search; Domain Ontology; Metadata Repository.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ANEXOS

1. Introdução

Este capítulo descreve o problema de pesquisa, sua relevância, a hipótese de solução e os objetivos da pesquisa, além de justificar a escolha dos métodos de avaliação da proposta e apresentar a organização do documento. A seção 1.1 fornece uma visão geral sobre Integração de Informações e nesta seção é enunciado o problema de pesquisa e a hipótese de solução identificada para este problema. Em seguida, o objetivo do trabalho é exposto na seção 1.2, acompanhado da proposta de solução. Os métodos de pesquisa utilizados são elencados na seção 1.3 juntamente com a justificativa para a sua escolha. Por fim, a seção 1.4 descreve a organização da dissertação.

1.1 Contexto e Motivação

Integração de informações é um processo crítico em instituições que possuem uma infinidade de fontes de informações, assim como para o progresso em larga escala de projetos científicos onde os dados são produzidos de forma independente por vários pesquisadores e para melhorar a cooperação entre os órgãos das estruturas governamentais através da transferência de dados durante a prestação de serviços públicos (Halevy et. al. 2006). Na visão da Web Semântica, os dados serão manipuláveis por máquinas e permitirão a geração e integração de informações através da interpretação semântica dos símbolos, aumentando o potencial de humanos e máquinas trabalharem em cooperação (Berners-Lee *et. al.* 2001), (Nigel *et. al.* 2006). Os desafios encontrados neste contexto em relação à interpretação semântica e os silos de dados são versões em escala Web do que as empresas vêm enfrentando por anos em suas demandas por integração de informações (PricewaterhouseCoopers 2009), (Bernstein e Haas 2008), (Halevy 2008).

As iniciativas de mapeamento da Arquitetura de Informação (Botto 2004) têm por objetivo garantir a Governança das Informações na organização tanto no que diz respeito à identificação de fontes de informação necessárias ao negócio, dentro e fora da organização, quanto à reutilização e integração das fontes existentes. A

necessidade de combinar informações a partir de múltiplas fontes surge com freqüência, em função das constantes mudanças nas organizações causadas por fusões e aquisições de empresas, reestruturações internas, interação com parceiros de negócio e disseminação de informações para aumentar a transparência organizacional, entre outras (Halevy 2008).

Após o levantamento dos requisitos de uma nova necessidade de informação, deve ser realizada uma busca por fontes de fontes de informação. O resultado desta busca irá determinar se a fonte de informação que atende a esta necessidade já existe dentro ou fora da organização e pode ser reutilizada ou se uma nova fonte deve ser desenvolvida. Se esta necessidade envolver o acesso, combinação e apresentação de informações de mais de uma fonte então a busca é a primeira etapa de um processo de integração de informações.

Integração de informações é um processo interativo e iterativo. Na maioria das organizações, diferentes fontes de informação coexistem e se sobrepõem em seu conteúdo, exigindo conhecimento de domínio para descobrir, compreender e integrar informações. Em geral, os sistemas de informação existentes não foram projetados para serem integrados e suas respectivas fontes de informação já disponíveis foram modeladas para atender a um contexto específico (Alexiev *et. al.* 2005), (Ziegler e Dittrich 2007). Em função do crescimento em quantidade e diversidade dos repositórios de informações, a busca por fontes de informação e a identificação do relacionamento entre elas se tornou uma atividade laboriosa.

A construção de um sistema de integração de informações (SII) através de um processo de integração de informações visa proporcionar acesso uniforme a um conjunto de fontes de informações heterogêneas e autônomas e fornecer uma visão integrada das informações. Um SII abstrai a complexidade de localização e acesso às fontes de informação além de resolver os conflitos entre as fontes para realizar a fusão dos dados. Várias abordagens foram propostas para construção de um SII, dando ênfase às soluções dos problemas estruturais e técnicos (Ziegler e Dittrich 2007), mas a integração semântica de informações deve garantir que somente dados relacionados com o mesmo conceito, ou pelo menos com conceitos semanticamente similares, serão combinados, tornando necessário tratar aspectos semânticos durante a integração de informações.

Esta proliferação de fontes de informação implica que uma necessidade de informação poder ser atendida por uma variedade de recursos disponíveis que armazenam dados sobre o mesmo domínio e possuem características de qualidade distintas. Para que um mecanismo de busca identifique estes recursos, é necessário que seu conteúdo esteja descrito de modo explícito e preciso através de metadados.

Além de encontrar os recursos que contêm a informação pertinente, também é necessário decidir entre diferentes recursos que têm as mesmas informações aquele que melhor atende a uma demanda específica (Cui *et. al.* 2001).

Conforme definido na NISO (2004) "*metadados são informações estruturadas que descrevem, explicam, localizam e tornam mais fácil a recuperação, o uso e o gerenciamento de um recurso de informação*". Os metadados podem ser armazenados junto ao recurso que descrevem ou dentro de um repositório para este fim específico, chamado de repositório de metadados. Este é um componente do sistema de informação para registro de metadados que fornece uma camada de abstração para descrever, gerenciar e consultar metadados de forma sistemática em ambientes distribuídos em grande escala. Alguns exemplos de sistemas de registro de metadados disponíveis para consulta na Web são: METeOR[1] (*Metadata Online Registry*) do governo da Austrália para informações sobre saúde, habitação e serviços, NBII[2] (*National Biological Information Infrastructure - Metadata Clearinghouse*) e EDR[3] (*Environmental Data Registry*) da Agência de Proteção Ambiental, sendo estes dois dos Estados Unidos, o Sistema de Metadados do IBGE[4], que descreve o acervo de informações públicas nas áreas de Estatística e de Geografia e o Catálogo de Metadados da INDE[5] (Infraestrutura Nacional de Dados Espaciais).

A maioria dos esquemas de metadados é composta por elementos contendo descrições em linguagem natural, o que permite o uso e interpretação humana do seu conteúdo e a aplicação de buscas por palavras-chave em repositórios de metadados. Esta busca visa obter um panorama das fontes disponíveis (Bernstein e Haas 2008) e serve como ponto de partida para exploração de dados. A catalogação de fontes de informação em um repositório de metadados pode aumentar o seu potencial de reutilização e integração. Porém, metadados descritivos não são suficientes para permitir a descoberta consistente e inequívoca das fontes de informação que contêm os dados necessários para atender a uma necessidade de informação e do relacionamento entre elas.

A partir deste contexto, o **problema** de pesquisa identificado é que, assim como na busca por documentos, a busca por fontes de informação em um repositório de metadados pode apresentar baixa precisão (número de fontes de informações

[1] http://meteor.aihw.gov.au

[2] http://search.nbii.gov/

[3] http://www.epa.gov/edr/

[4] http://www.metadados.ibge.gov.br/

[5] http://www.metadados.inde.gov.br/geonetwork/srv/br/main.home

relevantes recuperadas / número de fontes de informações recuperadas) e baixa cobertura (número de fontes de informações relevantes recuperadas / número de fontes de informações relevantes existentes na organização) do resultado. Isto pode acontecer em função de:

(1) divergência entre os termos da consulta e os utilizados no conteúdo dos elementos descritivos do esquema de metadados;

(2) conflitos, de nomenclatura e estrutura, devido aos diferentes contextos em que as fontes de informação foram modeladas;

(3) perda semântica dos processos de modelagem, desenvolvimento e catalogação das fontes de informação e

(4) nem todas as fontes de informação relevantes existentes na organização estão registradas no repositório de metadados.

Para aumentar a eficiência de ferramentas de recuperação de documentos, ontologias e mecanismos de inferência são utilizados para explorar o conhecimento do domínio e compartilhar a mesma estrutura de informação entre pessoas e agentes de software no processo de recuperação de documentos. Esta abordagem é chamada de busca semântica (Mangold 2007)**.** A atividade de busca por fontes de informação também depende do conhecimento do domínio e sua eficiência está relacionada com a completude e a capacidade de exploração do repositório de metadados onde as mesmas estão catalogadas.

Considerando este cenário, a **hipótese** levantada é que a utilização de busca semântica assistida pelo usuário irá tornar o resultado da atividade de busca por fontes de informação mais eficiente, no que diz respeito à precisão e cobertura. A busca semântica permite minimizar os problemas decorrentes de divergência terminológica, conflitos das fontes de informação e perda semântica.

1.2 Objetivos

O **objetivo** desta pesquisa acadêmica é tornar a busca por fontes de informação em repositórios de metadados mais eficiente, no que diz respeito à precisão e cobertura dos resultados, facilitando o processo de integração de informações e aumentando o potencial de reuso das mesmas. Para isto é necessário identificar a semântica da consulta e direcioná-la às fontes relevantes desta coleção, além de tornar explícita a semântica destas fontes e a similaridade semântica entre elas.

A proposta de solução para este problema é uma arquitetura lógica suportada por busca semântica para recuperação de fontes de informação em repositórios de

metadados. Ontologias de domínio são utilizadas para formalizar as relações semânticas entre os conceitos de um domínio particular, para recuperar o melhor conjunto de fontes de informação semanticamente similares e aumentar a semântica de metadados descritivos através de anotações semânticas.

Esta arquitetura propõe uma extensão das funcionalidades de busca e catalogação de fontes de informação no repositório de metadados. A busca explora o conhecimento do domínio e permite que o usuário especifique os conceitos envolvidos em sua necessidade de informação de uma forma mais explícita, precisa e baseada em sua intenção de busca. A anotação semântica, realizada em conjunto com a catalogação das novas fontes de informação, permite identificar em quantas e quais fontes de informação os dados de um determinado conceito estão contidos e quais os conceitos presentes em uma fonte de informação específica. Esta característica atende à integração semântica de informação, pois através da ontologia de domínio é possível identificar a similaridade semântica entre os conceitos. Além disto, o rastreamento das consultas fornece insumos para evolução da ontologia de domínio e priorização da anotação semântica de fontes de informação legadas.

A arquitetura proposta pode ser usada no cenário hipotético descrito a seguir.

Uma empresa possui diversos sistemas de informação que foram projetados para propósitos específicos, por profissionais diferentes, desenvolvidos usando tecnologias distintas e que individualmente atendem plenamente aos requisitos de seus usuários operacionais. Com o objetivo de facilitar o acesso aos dados de todos os sistemas desta instituição, a equipe responsável por cada sistema desenvolveu um conjunto de serviços para encapsular as particularidades tecnológicas e a complexidade de seu modelo de dados. Estes serviços e seus respectivos sistemas foram catalogados por seus desenvolvedores em um repositório de metadados onde os demais desenvolvedores podem, através de uma busca com palavras-chave, encontrar o serviço que recupere um conjunto de dados de interesse.

Nesta organização surge uma nova necessidade de informação. O gerente de cada filial deseja monitorar as informações dos clientes depois que estes realizam uma compra em sua filial. Para atender a esta necessidade de informação, um desenvolvedor irá construir um portal para visualização destes dados, mas antes ele precisa descobrir quais são os serviços que serão utilizados para fornecer os dados. Ele inicia uma busca no repositório de metadados usando a palavra chave “**cliente**”.

Se esta busca não fosse suportada pela definição semântica dos conceitos do negócio, seriam recuperados três serviços para obtenção de dados dos sistemas que são usados pelos departamentos de faturamento, vendas e suporte técnico. Mas através da busca semântica, são identificados na ontologia de domínio três conceitos

associados à palavra "**cliente**": "**Cliente** Faturado", "**Cliente** em Prospecção" e "**Cliente** com Garantia". O desenvolvedor analisa as opções e entende que o "Cliente em Prospecção" é aquele que ainda está em processo de negociação e não realizou uma compra e por isso seleciona os conceitos "Cliente Faturado" e "Cliente com Garantia". Somente dois serviços são recuperados, o primeiro obtém dados de "Cliente Faturado" do sistema de faturamento e o segundo de "Cliente com Garantia" do sistema de suporte técnico.

1.3 Métodos de Pesquisa

Dois métodos, experimento e estudo de caso, foram utilizados na avaliação da proposta para confirmar se a hipótese levantada produz os resultados esperados.

O experimento compara a precisão e a cobertura das consultas realizadas no repositório de metadados com o apoio da arquitetura proposta em relação a outras abordagens existentes na literatura. Este método foi escolhido por permitir a análise quantitativa das variáveis dependentes (precisão, cobertura e medida F) que determinam o desempenho da abordagem a partir da manipulação de variáveis independentes (consultas executadas). O resultado obtido foi que a busca semântica atingiu melhor desempenho uma vez que esta explora o conhecimento do domínio e permite que o usuário especifique de maneira mais precisa a sua intenção durante a atividade de busca por fontes de informação. Ao final do experimento também foi aplicado um questionário aos participantes para coletar informações quanto à utilização da ferramenta, o perfil dos mesmos e do ambiente organizacional onde estes estão inseridos.

O segundo método de avaliação (estudo de caso) buscou evidenciar a viabilidade, benefícios e dificuldades de aplicação da proposta em um ambiente real. Este método foi escolhido por permitir investigar um fenômeno contemporâneo dentro de seu contexto na vida real (Yin 2005) e analisar a aplicação de sistemas de informação em organizações considerando as práticas organizacionais e o conhecimento das pessoas envolvidas para entender e explicar um fenômeno social (Myers 1997). O Catálogo de Informações do portal *DadosGov* COI-PR foi utilizado nesta fase por ser uma iniciativa do governo federal brasileiro em disponibilizar dados em formatos abertos para serem reutilizados e integrados com outras informações pelo próprio governo e pela sociedade em geral. Uma cópia do repositório de metadados, além de documentos, modelos, especificações e acesso a área restrita do portal, foram cedidos pelo SERPRO. Reuniões presenciais e comunicações por e-mail com a equipe responsável pela aplicação foram realizadas de modo a coletar

informações adicionais, dirimir as dúvidas quanto aos documentos e apresentar a arquitetura da proposta.

1.4 Organização da Dissertação

Com o objetivo de apresentar o desenvolvimento da pesquisa acadêmica, esta dissertação foi organizada da seguinte forma:

O presente capítulo descreveu o problema e sua relevância, a hipótese de solução e os objetivos da pesquisa, além de justificar a escolha dos métodos de avaliação da proposta.

O capítulo 2 apresenta a fundamentação teórica do contexto do problema e da abordagem de solução. Este capítulo descreve os principais conceitos envolvidos e abordagens adotadas para Integração de Informações, dando destaque à perspectiva semântica e ao papel dos modelos de dados que compõem a Arquitetura de Informações de uma Organização, em especial das Ontologias e Metadados na atividade de descoberta de fontes de informação organizacionais.

No capítulo 3 são descritos os componentes de uma arquitetura lógica suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados. Essa arquitetura faz uso de ontologias de domínio para formalizar as relações semânticas entre os conceitos de um domínio particular, para recuperar o melhor conjunto de fontes de informação semanticamente similares e aumentar a semântica de metadados descritivos armazenados no repositório de metadados através do uso de anotações semânticas.

No capítulo 4 é apresentado o protótipo da arquitetura implementado para o experimento.

O capítulo 5 descreve o experimento realizado como a primeira fase de avaliação da proposta, detalhando a dinâmica do experimento, seus resultados e a análise das respostas do questionário enviado aos participantes do experimento.

A segunda fase de avaliação, um estudo de caso utilizando o Catálogo de Informações do portal *DadosGov* COI-PR, é abordada no capítulo 6. Um exemplo de busca por fontes de informação com apoio da arquitetura é detalhado assim como o resultado das buscas realizadas através de uma aplicação, disponibilizada na internet, como prova de conceito para o estudo de caso.

O capítulo 7 conclui a dissertação, comparando a proposta com trabalhos relacionados que tratam de integração semântica de informações e consolidando os resultados obtidos em ambos os métodos de avaliação, além de ressaltar as principais contribuições e limitações da pesquisa e sugerir trabalhos futuros.

2. Fundamentação Teórica

A Web Semântica é descrita como a evolução da web atual, que consiste de uma enorme coleção de documentos interligados, para uma web onde dados são manipuláveis por máquinas e permitem a geração de informações através da interpretação semântica dos símbolos, aumentando o potencial de humanos e máquinas trabalharem em cooperação (Berners-Lee *et. al.* 2001), (Nigel *et. al.* 2006). Mas os desafios encontrados na Web em relação à semântica dos dados e os silos de dados são versões em escala Web do que as empresas vêm enfrentando por anos em suas demandas por integração de informações (PricewaterhouseCoopers 2009), (Bernstein e Haas 2008), (Halevy 2008). Em função do crescimento em quantidade e diversidade dos repositórios de informações nas organizações, a descoberta de fontes de informação e do relacionamento entre elas se tornou uma atividade laboriosa.

Neste cenário, a integração de informações depende do conhecimento de onde a informação está armazenada e de como acessar estas fontes além do entendimento de sua organização lógica para relacioná-las e transformá-las em uma visão integrada a ser apresentada (Patrick 2005). Essa visão lógica única deve garantir que somente dados relacionados com o mesmo conceito, ou pelo menos com conceitos semanticamente similares, serão combinados (Ziegler e Dittrich 2007), tornando necessário tratar aspectos semânticos durante a integração de informações.

Este capítulo descreve os principais conceitos envolvidos e abordagens adotadas para Integração de Informações, dando destaque à perspectiva semântica e ao papel dos modelos de dados que compõem a Arquitetura de Informações de uma Organização, em especial das Ontologias e Metadados na atividade de descoberta de fontes de informação organizacionais. As ontologias como estruturas conceituais e lógicas permitem a organização dos metadados de acordo com seus princípios semânticos de modo que possam ser processados por máquinas e compreendidos por humanos, enquanto os metadados descritivos transformam dado em informação por descreverem o que o dado representa dentro de seu contexto e auxiliam na busca por fontes de informação.

2.1 Arquitetura de Informação

A Arquitetura Corporativa, mais recentemente denominada Arquitetura Empresarial, é composta por modelos descritivos que definem o negócio, a informação e a tecnologia subjacente que permitem o alinhamento de TI com o negócio. A Arquitetura de Informação, como parte da Arquitetura Corporativa (Botto 2004), tem por objetivo garantir a Governança das Informações na organização tanto no que diz respeito à identificação de fontes de informação necessárias ao negócio, dentro e fora da organização, quanto à reutilização e integração das fontes existentes.

A Governança da Informação é realizada através de processos de gestão do acervo de informações apoiados por tecnologia ao longo de todo o ciclo de vida. Estes processos lidam com atividades desde a concepção das fontes de informação até o rastreamento da produção e do uso das mesmas. Para isto, é necessária uma mudança na cultura organizacional de modo a reconhecer a informação como um ativo para a organização e a criação de novos papéis como os arquitetos de informação e os tutores das informações (*Information Stewards*) (Godinez *et. al.* 2010).

Os modelos de dados que compõem a Arquitetura de Informação de uma Organização representam o seu acervo de dados em diferentes perspectivas e níveis de abstração, e são usados principalmente como artefatos de comunicação. Um modelo de alto nível representa os conceitos do negócio e seus relacionamentos usando descrições concisas e viabiliza o consenso no que diz respeito à terminologia e definições. Os conceitos, relacionamentos e atributos são representados através de termos que têm significado para o negócio e que respondem a questões do tipo *"Quem", "O quê", "Onde", "Quando", "Por que" e "Como"* (Hoberman *et. al.* 2009). Neste contexto, ontologias podem ser consideradas como modelos conceituais de dados que, junto com outros artefatos, permitem obter uma visão das informações corporativas como um todo (Azevedo *et. al.* 2009).

Suponha um cenário onde: para o departamento de faturamento, a definição de Cliente seja "*alguém que comprou um produto da empresa e para o qual tenha sido emitida uma fatura*"; para o departamento de vendas seja "*todo aquele, pessoa física ou jurídica, que está em processo de negociação para compra de um produto*"; e para o departamento de suporte ao Cliente seja "*quem comprou um produto que está dentro da garantia*". Através do processo de construção de um modelo de alto nível, como o apresentado na figura 2.1, é possível evidenciar a existência de diferentes perspectivas sobre o que é um Cliente dentro de uma organização, onde as definições anteriores para cliente foram representadas respectivamente pelos conceitos "Cliente Faturado", "Cliente em Prospecção" e "Cliente com Garantia". Estas definições não

são consideradas conflitantes, no entanto não haviam sido salientadas quando os sistemas de informação que atendem as funções departamentais foram construídos isoladamente. Mesmo assim é necessário torná-las explicitas quando se deseja uma visão integrada destas informações como, por exemplo, em sistemas de *Master Data Management* (MDM), onde são consolidadas as informações chaves para o negócio.

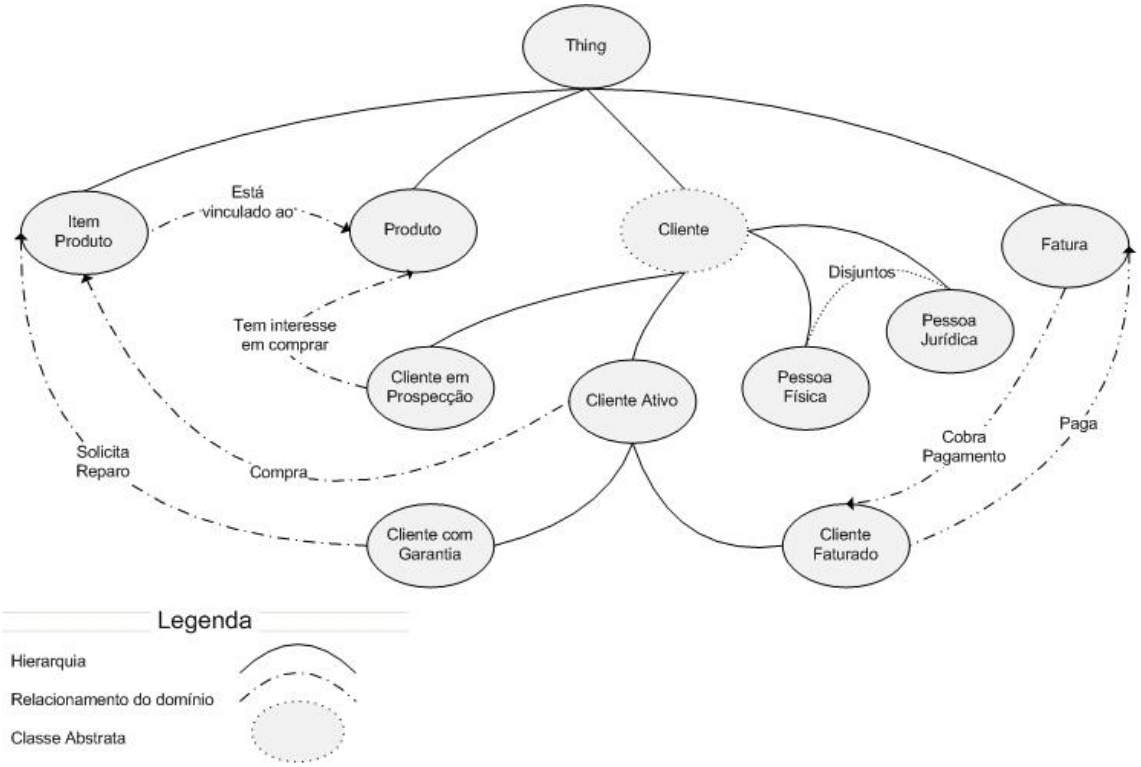


Figura 2.1 Fragmento de um Modelo de Alto Nível

Os modelos de dados são importantes instrumentos para que usuários e arquitetos de informação obtenham um entendimento sobre como gerenciar e descobrir novas oportunidades de uso da informação como recurso corporativo. Este entendimento requer muito mais que definições sobre os conceitos e a estrutura de suas fontes de informação. Orientações sobre como utilizar e interpretar o acervo de dados disponível, descritas na linguagem dos usuários, permitem a definição de um mapa dos itens de informação relevantes para o negócio (Evernden e Evernden 2003).

2.1.1 Metadados

Conforme definido pela NISO (2004), "*metadados são informações estruturadas que descrevem, explicam, localizam e tornam mais fácil a recuperação, o uso e o gerenciamento de um recurso de informação*". Os metadados podem ser classificados em três tipos: (1) metadados descritivos descrevem uma fonte de informação para fins de identificação e recuperação utilizando elementos como título,

autor, resumo e palavras-chave, (2) metadados estruturados descrevem a organização interna dos objetos e das relações entre eles; o exemplo mais comum é o esquema do banco de dados, e (3) metadados administrativos apóiam as atividades de gerenciamento do acervo de recursos de informação como controle de permissões de acesso, localização de arquivos e critérios de avaliação da qualidade.

Metadados descritivos revelam parte da semântica das fontes de informação, como a intenção de uso, além da descrição ou resumo do seu conteúdo (Sheth 1998), fornecendo o contexto para entendimento dos dados através do tempo. Contexto é um aspecto importante na semântica dos dados e ajuda os analistas a interpretarem o significado atribuído a estes dados (Agrawal *et. al.* 2009). Ontologias também permitem atribuir significado aos dados uma vez que estas fornecem um vocabulário explícito e compartilhado para identificação de conceitos, seus atributos e do relacionamento entre estes.

As informações sobre qualidade de dados associadas às fontes de informação são uma parte do grupo de elementos de metadados administrativos. Qualidade de dados é um conceito multidimensional que representa a visão, os critérios e as métricas para avaliar, interpretar e melhorar as fontes de informação além de permitir a seleção das dimensões relevantes dentro de um contexto de aplicação (Sattler 2008). As medidas são qualitativas, baseadas no julgamento de especialistas e usuários, e quantitativas, considerando as dimensões de avaliação (Batini e Scannapieco 2006).

Para aumentar a confiabilidade dos usuários em relação às fontes de informações consideradas relevantes, além do conhecimento quanto à qualidade das mesmas, é necessário ter ciência de que estas são governadas por processos adequados, são gerenciadas por seus tutores (*Information Steward*) e possuem controle de acesso adequado, e para isto estas informações também devem ser registradas como metadados (Godinez *et. al.* 2010).

De acordo com Godinez *et. al.* (2010) metadados também podem ser classificados em técnicos e de negócio. Metadados técnicos contém informações referente a localização, protocolos de acesso e esquema físico e lógico das fontes de dados. Enquanto que, a maioria das informações textuais sobre os dados que usuários de negócio acrescentam nos modelos de dados de alto nível (Hoberman *et. al.* 2009) e fornecem esclarecimento contextual são consideradas metadados de negócio como, por exemplo, as definições dos conceitos no negócio, o responsável pela definição, indicadores de qualidade das informações, critérios de segurança, indicação dos produtores e dos consumidores da informação além de vocabulários e taxonomias usadas para definir a terminologia e hierarquia dos conceitos do negócio.

Metadados podem ser extraídos de diversas fontes como políticas e procedimentos organizacionais, modelos de processos de negócio, mas principalmente de pessoas, no papel de responsáveis pelas aplicações e processos que geram e manipulam as informações (Godinez *et. al.* 2010).

2.1.1.1 Padrões de Metadados

Para facilitar o compartilhamento de informações entre instituições produtoras e consumidoras e entre ferramentas, foram desenvolvidos padrões que tratam do conteúdo, intercâmbio eletrônico e de modelos de dados para gerenciamento de metadados. Estes padrões são especificados através de esquemas de metadados, compostos por elementos concebidos para fins específicos cuja definição fornece a semântica do esquema e, opcionalmente, por regras que determinam como o conteúdo dos elementos deve ser preenchido como, por exemplo, lista de valores permitidos (NISO 2004).

Apesar do esforço empregado no desenvolvimento e adoção de padrões de metadados dentro de uma comunidade de usuários, a maioria destes padrões sofre modificações através de extensões e perfis. Extensões são criadas para adicionar elementos que suportam as necessidades específicas de um determinado grupo. Os perfis restringem o número de elementos opcionais e aperfeiçoam definições do conteúdo de elementos.

Dois esquemas de metadados muito utilizados para descrição e recuperação de recursos de informação são: Dublin Core e ISO 19115:2003.

O objetivo inicial do padrão Dublin Core (DC) foi definir um conjunto mínimo de elementos que pudessem ser usados pelos produtores de informação na *Web* para descrever os seus próprios recursos de modo simples e conciso. Este padrão é composto por 15 elementos, todos opcionais e que podem ser repetidos. Quando associados a qualificadores permitem que a semântica dos elementos seja refinada e que uma lista de valores permitidos ou formato seja especificado (DCMI 2010).

Este padrão foi transformado nas normas ANSI/NISO Z39.50 e ISO 15836:2003. A sua simplicidade foi o principal fator para alavancar a sua ampla utilização, porém também pode ser considerado como um problema uma vez que não permite especificar a semântica de seus elementos de modo mais expressivo (Breitman *et. al.* 2007).

O padrão internacional ISO 19115:2003 especifica um esquema de metadados para descrição de objetos geográficos (Breitman *et. al.* 2007), adotado como padrão para exportação de esquemas de catálogos geográficos. Este padrão fornece um conjunto de elementos essenciais e outros opcionais, e cada aplicação pode definir um

perfil que contém a lista de elementos opcionais utilizados. Este padrão vem sendo amplamente utilizado por iniciativas de Infraestrutura de Dados Espaciais (IDE) na criação de seus perfis como é o caso do MIG (Portugal) e NEM (Espanha) da iniciativa européia chamada INSPIRE (*Infrastructure for Spatial Information in Europe*) (Soares et. al. 2010), entre outros.

O objetivo da IDE é fornecer uma coleção integrada de serviços que permita aos usuários descobrir e acessar dados geográficos de diversas fontes e para isto depende da definição de um esquema de metadados que suporte a publicação destas fontes de informação assim como o acesso ao conteúdo das mesmas. No Brasil, a Infraestrutura Nacional de Dados Espaciais (INDE) definiu a partir do ISO 19115:2003 uma versão completa para o seu perfil de metadados, chamado de Metadados Geográficos Brasileiro (MGB), com 82 elementos, divididos em 10 seções, e uma versão sumarizada com 21 elementos (CONCAR 2009).

2.1.1.2 Repositório de Metadados

Os metadados podem ser armazenados junto ao recurso que descrevem ou em um repositório para este fim específico. Um repositório de metadados é um componente do sistema de informação para registro de metadados e contém esquemas de metadados, perfis, listas de codificação de valores e do conteúdo dos elementos, mas não contém os dados em si. Este sistema fornece uma camada de abstração para descrever, gerenciar e consultar metadados de forma sistemática em ambientes distribuídos em grande escala.

Repositórios de modelos de dados permitem o armazenamento de metadados capturados nas etapas de modelagem como definições de conceitos junto com os respectivos elementos dos modelos. De modo que seja possível compartilhar estes metadados e outros metadados capturados através de atividades como modelagem de processos e desenvolvimento de aplicações é necessário um repositório corporativo de metadados (Hoberman *et. al.* 2009) para registrar informações sobre os processos que criam, usam ou atualizam as fontes de informação assim como os componentes de hardware e software que hospedam esses processos.

A capacidade de gerenciamento de metadados atende a alguns requisitos para o estabelecimento da Arquitetura de Informações como: (1) a criação e manutenção de um vocabulário compartilhado de termos do negócio, o que facilita a comunicação entre os profissionais de TI e os usuários, (2) a identificação da qualidade dos dados por parte de usuários técnicos e de negócio, o que permite estabelecer o nível de confiança nas fontes de informação, e (3) a avaliação do impacto de mudanças dos

requisitos de negócio e da infraestrutura técnica de TI nas fontes de informação e nos processos que elas suportam (Godinez *et. al.* 2010).

Por isso a eficiência no gerenciamento de metadados não é um objetivo final por si só, mas sim um recurso para alavancar a Governança de Informações. Se o repositório de metadados for explorado por uma ferramenta de busca que torne eficiente a descoberta de fontes de informação, isto aumentaria a reutilização e reduziria a heterogeneidade semântica das fontes, ao fazer com que gerentes, arquitetos e desenvolvedores de sistemas se tornem mais conscientes sobre a existência das mesmas na organização (Smith *et. al.* 2009).

2.1.2 Integração de Informações

As fontes de informação dentro das organizações vão desde aplicações empacotadas como *Customer Relationship Management* (CRM) até sistemas de arquivos contendo documentos, planilhas, apresentações, imagens e vídeos, passando por bases de dados armazenadas em Sistemas Gerenciadores de Bancos de Dados (SGBD).

A integração de informações permite combinar informações de diferentes fontes de informação de modo a criar uma visão integrada que facilite o acesso e o reuso das mesmas. A construção de um sistema de integração de informações (SII) proporciona acesso uniforme a um conjunto de fontes de informações heterogêneas e autônomas, abstraindo a complexidade de acesso às fontes, para fornecer como resposta esta visão integrada. Várias abordagens foram propostas e construídas com diferenças no nível de abstração em que a integração ocorre, dando ênfase às soluções dos problemas de integração de informações do ponto de vista estrutural e técnico. O esquema de classificação, proposto por Ziegler e Ditrich (2007), divide as soluções em seis categorias de acordo com uma perspectiva arquitetural: integração manual, interface comum de usuário, integração por aplicação, integração por *middleware*, acesso uniforme aos dados e armazenamento comum de dados.

Quando os usuários são responsáveis por localizar, acessar, entender e integrar os dados a integração é manual, porém se os dados são apresentados separadamente em uma mesma interface e o usuário é quem realiza a integração fica caracterizada uma abordagem de interface comum de usuário, uma vez que somente a localização e o acesso as fontes de dados isoladamente foi realizada pelo sistema. Um portal desenvolvido com a finalidade de fornecer conteúdo dinâmico, que exibe informações sobre um determinado assunto sem realizar nenhuma operação que inter-relacione os dados (Bernstein e Haas 2008), é um exemplo de solução de integração de informações através de interface comum de usuário.

As demandas por integração de informações também podem ser atendidas pelo desenvolvimento pontual de aplicações que acessam um conjunto pré-determinado de fontes de informações, realizam tratamento em relação aos dados recuperados e apresentam as informações integradas aos usuários. Essas abordagens são eficientes somente quando o número de fontes, interfaces de acesso e formatos de dados é pequeno e por isso não tem escalabilidade. Para reduzir a complexidade de acesso aos diversos formatos das fontes de dados é possível utilizar um *middleware*, mas a lógica da aplicação ainda precisa realizar a integração para apresentar o resultado.

Nas abordagens de acesso uniforme aos dados, como sistemas de integração virtual de informações baseados em mediadores de consulta e federação de bancos de dados, as aplicações possuem uma visão global e unificada de fontes de dados, mas estas se mantêm fisicamente distribuídas e autônomas. As fontes de dados podem variar de bancos de dados e sistemas legados até formulários, serviços web e arquivos texto com formato semi ou não estruturados que são acessados através de *wrappers*. Os usuários e aplicações submetem as consultas ao esquema mediado, que é um esquema lógico que contém somente os aspectos de domínio relevantes para a aplicação. Através dos mapeamentos existentes entre os esquemas das fontes de dados locais e o esquema mediado, a consulta original é transformada em consultas a serem enviadas aos *wrappers* (Halevy 2008).

O armazenamento comum de dados é identificado nas soluções de integração de informações onde os dados são extraídos de suas fontes de dados originais e armazenados em um repositório central único. *Data Warehouse* (DW) é um exemplo por ser uma solução para integração materializada de informações que se caracteriza por ser "*uma coleção de dados orientados por assunto, integrados, não voláteis, variável em relação ao tempo, de apoio às decisões gerenciais*" (Inmon 1997). Nesta solução, os dados de sistemas de informação transacionais são extraídos, transformados e carregados em um banco de dados centralizado para realização de processamento analítico on-line (OLAP) de informações integradas e agregadas.

De modo geral, a maior parte dos sistemas de informação não é projetada inicialmente para ser integrada, trazendo grandes desafios quando isso se faz necessário (Alexiev *et. al.* 2005). Essa característica também se reflete nas fontes de informação destes sistemas uma vez que seus esquemas são projetados por diferentes pessoas para diferentes aplicações e em função disto são encontradas diferenças entre eles, mesmo quando pertencem ao mesmo domínio (Halevy 2008). Os conflitos gerados por essas diferenças foram definidos por Batini *et. al.* (1986) como "*duas (ou mais) representações não idênticas do mesmo conceito*" (intensão; do

Inglês *intension* [1]), porém também podem ocorrer no nível das instâncias (extensão; do Inglês *extension* [2]).

De acordo com Bleiholder e Naumann (2008), diferentes formas de representação do mesmo objeto do mundo real nas fontes de dados podem resultar em três tipos de conflitos: (1) conflitos esquemáticos, por exemplo, quando o mesmo atributo possui nomes, tipos ou restrições de integridade diferentes, (2) conflitos de identidade, nos casos onde a forma de identificação do mesmo objeto do mundo real é diferente de uma fonte em relação à outra e (3) conflitos de dados, se os valores dos atributos semanticamente equivalentes forem diferentes de uma fonte para outra. Os conflitos de dados ainda podem ser divididos em incerteza, nos casos em que o valor é desconhecido em uma das fontes, e contradição, se a mesma propriedade para o mesmo objeto possui valores distintos em duas ou mais fontes.

2.1.2.1 Processo de Integração de Informações

A construção de sistemas de integração de informações requer um processo interativo para localização das fontes de informação, das semelhanças e dos conflitos entre elas e a definição do tratamento destes conflitos (Bleiholder e Naumann 2008). O esquema mediado pode ser gerado através de um processo de integração de esquemas ou pode ser definido antes da análise das fontes de dados, a partir de uma necessidade de informação. Neste último, a origem de seus dados é identificada através de um processo de mapeamento de esquemas.

O processo de integração de esquemas, conforme proposto em Batini *et. al.* (1986), é composto por quatro atividades: pré-integração, comparação de esquemas, conformação de esquemas e fusão e reestruturação. Na atividade de pré-integração é realizada a análise das fontes de dados disponíveis para definir a política de integração a ser seguida. Nesta análise, dependendo da quantidade e complexidade das fontes de dados existentes, é determinada a precedência entre as fontes a serem submetidas às próximas etapas em cada iteração. Na comparação de esquemas são determinadas as correspondências e os conflitos entre os conceitos dos esquemas selecionados através de operações de casamento de esquemas. A etapa de conformação de esquemas tem por objetivo resolver os conflitos sintáticos, estruturais e semânticos que existem entre os esquemas. O processo finaliza com as etapas de fusão e reestruturação quando as instâncias são combinadas e os esquemas são reestruturados de modo a gerar um esquema integrado que seja mínimo, completo e

[1] The sum of the attributes contained in a term. The collective attributes, qualities, or marks that make up a complex general notion; the comprehension, content, or connotation; -- opposed to extension, extent, or sphere.

[2] The class of objects designated by a specific term or concept; denotation. Capacity of a concept or general term to include a greater or smaller number of objects; -- correlative of intension.

compreensível (Py *et. al.* 2009), (Mello e Heuser 2005). Este processo utiliza uma orientação *bottom-up*, pois parte das fontes de dados para construir o esquema mediado.

Quando o esquema mediado é definido a partir de uma necessidade de informação e antes da análise das fontes de dados, o primeiro passo é buscar as fontes que atendem a esta necessidade. As fontes recuperadas são submetidas a operações de casamento de esquemas para gerar os mapeamentos semânticos entre estas fontes e o esquema mediado. A etapa seguinte visa identificar as instâncias presentes em fontes de dados distintas que representam o mesmo objeto e o processo finaliza com a combinação destas instâncias. Este processo de integração segue uma orientação *top-down* (Mello 2002), (Bleiholder e Naumann 2008).

Cada iteração de um processo de integração de informações visa aumentar a completude e concisão das informações a serem fornecidas (Bleiholder e Naumann 2008). Completude significa que nenhuma informação será ignorada no resultado integrado à medida que novas fontes são incluídas a cada ciclo. A concisão mede a unicidade de representação de um objeto em um conjunto de dados. Para cada fonte incluída é necessário remover os atributos e dados redundantes ou mesclar dados e atributos duplicados em um único.

2.1.2.2 Casamento de Esquemas

A operação de casamento de esquemas (*schema matching*), utilizada tanto no processo de integração de esquemas quanto no processo de mapeamento de esquemas, permite a descoberta dos mapeamentos semânticos entre os elementos de cada esquema. Esta descoberta pode ser baseada nas instâncias, no nome e nas características dos elementos que compõem o esquema, na estrutura do esquema das fontes, em suas descrições e outras entradas auxiliares ou em uma combinação de métodos (Rahm e Bernstein 2001), (Halevy 2008). Na figura 2.2 é apresentada a classificação das abordagens de casamento de esquemas proposta por Rahm e Bernstein (2001). Estas abordagens são consideradas como *a posteriori* no sentido que seu objetivo é realizar o casamento de esquemas de fontes existentes.

Os mapeamentos candidatos gerados são associados a uma medida de similaridade estimada no intervalo de (0-1), permitindo o estabelecimento de um valor mínimo (*threshold*) para corte. Esses mapeamentos devem ser analisados por especialistas, pois apesar de uma parte da semântica das fontes de informação estar nestes insumos, esta não é descrita de forma suficientemente explícita e precisa para permitir a geração automática de resultados corretamente.

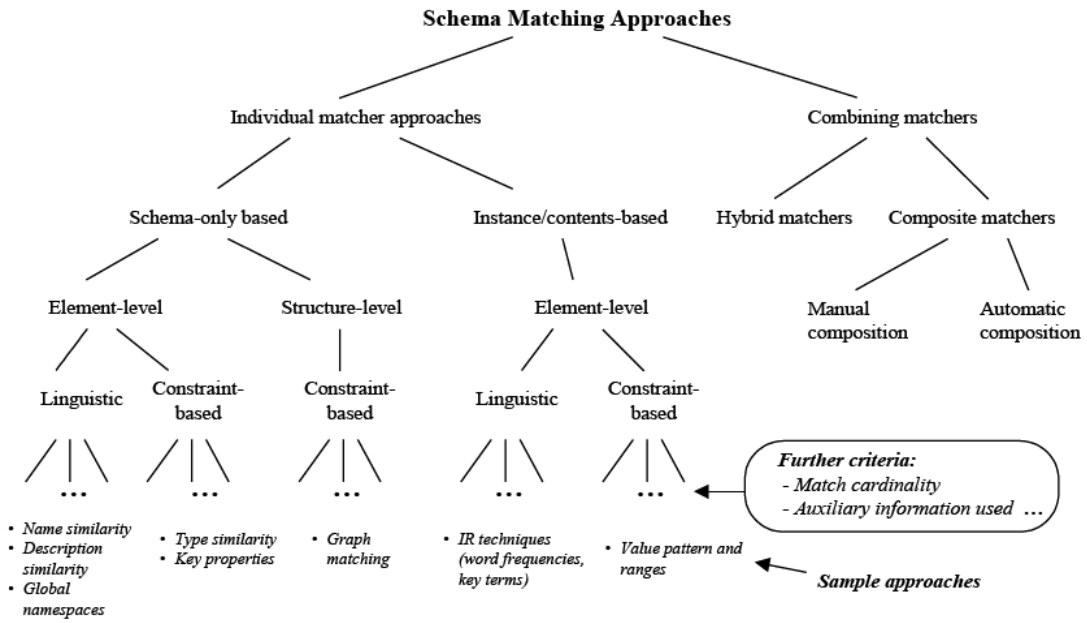


Figura 2.2 Abordagens de Casamento de Esquemas de acordo com (Rahm e Bernstein 2001).

As abordagens de casamento de esquemas *a posteriori* (denominadas sintática e semântica) analisadas por Casanova *et. al.* (2007) correspondem ao nível de esquema (ou intensão) e nível de dados (ou extensão) respectivamente da classificação proposta por Rahm e Bernstein (2001). O casamento de esquemas baseado no esquema pressupõe que a similaridade sintática implica em similaridade semântica, mas a existência de conflitos de esquemas geram mapeamentos incorretos. O casamento de esquemas baseado em instâncias depende da habilidade de detectar se duas instâncias em fontes de dados diferentes referem-se ao mesmo objeto do mundo real para resolver os conflitos de identidade.

Ortogonalmente também foi descrita uma terceira abordagem *a priori*. Nesta última, os profissionais que realizam a modelagem de dados nas organizações devem identificar um esquema padrão de dados adequado ou propor um novo e reutilizá-lo no desenvolvimento de novas fontes de informação. O uso de ontologias torna a abordagem *a priori* viável, uma vez que conceitos pertinentes ao domínio da fonte de informação que está sendo modelada podem ser extraídos, alinhados e publicados como uma ontologia de domínio.

A abordagem *a priori* torna a integração destas fontes mais fácil uma vez que evita a necessidade de formatação dos dados antes da integração, evita os conflitos de esquemas e a perda semântica uma vez que garante que o significado dos elementos é conhecido desde a concepção das fontes. O sucesso desta abordagem depende de uma metodologia para modelagem de dados que seja seguida por todas as equipes, sendo adequada em organizações e comunidades onde um processo de

Governança de Informações tenha sido estabelecido e adotado. Mas esta só é aplicável ao desenvolvimento de novas fontes de informação (e por isso chamada de *a priori*) e as abordagens *a posteriori* continuam sendo necessárias para integração de informações de fontes legadas.

A semântica também está incorporada em modelos conceituais, programas de aplicação e até nas mentes dos usuários e projetistas. Na modelagem de dados, a semântica é considerada como a interpretação dos projetistas e usuários sobre os requisitos de informação em um determinado contexto (Ziegler e Dittrich 2007). No que diz respeito aos modelos conceituais algumas propostas tem surgido para aumentar a sua qualidade semântica através do uso de ontologias de fundamentação (Castro *et. al.* 2010), (Guizzardi *et. al.* 2010) mas estas abordagens tem por objetivo apoiar a comunicação entre agentes humanos no uso destes artefatos e não a interpretação automática dos conceitos para integração das informações.

No nível dos modelos lógicos e físicos não é possível especificar exaustivamente a semântica associada aos dados e elementos de esquemas. Com isso as definições de esquemas não são semanticamente explícitas para permitir a interpretação dos dados de forma consistente e inequívoca, mas a semântica precisa e explícita das fontes de dados é essencial para gerar resultados integrados corretos e que façam sentido. O uso de uma ontologia de domínio, junto com um vocabulário compartilhado que referencia os conceitos contidos nas fontes de dados, permite minimizar os problemas decorrentes da heterogeneidade semântica (Ziegler e Dittrich 2007).

2.2 Ontologias

Na literatura podem ser encontradas várias definições para ontologias, uma definição muito citada é: "*Uma ontologia é uma especificação formal explicita de uma conceitualização compartilhada*" (Gruber 1993), (Fensel 2001), (Uschold e Gruninger 2004). Uma conceitualização é uma visão abstrata e simplificada do mundo que tem por propósito representar um fenômeno através dos conceitos relevantes e por isso uma ontologia é considerada um modelo. Nesta definição, explícita quer dizer que os tipos de conceitos presentes neste modelo e suas restrições de utilização são claramente definidos e compartilhados. Para ser compartilhada é necessário que a ontologia capture conhecimento consensual, isto é, esse conhecimento não deve ser restrito a alguns indivíduos, mas aceito por um grupo de pessoas.

Uma ontologia pode ser definida como um vocabulário específico dotado de significado para identificação de conceitos e dos relacionamentos entre estes de modo

a descrever determinados aspectos da realidade (Sheth 1998), combinando o entendimento humano de símbolos com a capacidade de processamento por máquinas. Sistemas que fazem uso de ontologias podem reconhecer ou entender o contexto de uma necessidade de informação e usá-las para limitar a sobrecarga de informações tanto através da elaboração de consultas mais precisas quanto pela filtragem e transformação da informação antes de apresentar ao usuário.

Conceitos ou classes são abstrações de um conjunto de objetos, lexicamente definidos por termos em linguagem natural (Giunchiglia e Zaihrayeu 2009) e semanticamente definidos por suas características e pelas relações com os demais conceitos. Além dos relacionamentos taxonômicos entre conceitos, outros relacionamentos ou propriedades de objetos podem ser representados assim como as características dos conceitos, através de atributos ou propriedades de dados. Indivíduos ou instâncias também podem fazer parte da ontologia.

2.2.1 Tipos de Ontologias

Guarino (1998) indica que ontologias devem ser construídas de acordo com o seu nível de generalidade seguindo a categorização representada na figura 2.3, onde as setas representam relações de especialização:

- **Ontologias de Alto Nível** (*Upper*) ou Genéricas ou de Fundamentação: descrevem conceitos gerais e abstratos ou meta-conceitos como espaço, tempo, matéria, objeto, evento, assunto, ação entre outros. Alguns exemplos são SUMO (*Suggested Upper Merged Ontology*) e OpenCyc, desenvolvidas pelo grupo de trabalho da IEEE, e DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*), desenvolvida pelo projeto *WonderWeb* de pesquisadores europeus. O objetivo destas ontologias é servir como base para criação de ontologias de domínio e tarefas e com isso permitir a integração entre as aplicações que fazem uso destas, pois se duas ontologias fazem referencia a um vocabulário comum então encontrar as correspondências entre elas se torna mais fácil (Noy 2004).
- **Ontologias de Domínio**: descrevem o vocabulário específico relacionado a domínios como Turismo, Geografia, Medicina e Biologia. Possuem o enfoque relacionado aos conceitos e objetos do universo de discurso.
- **Ontologias de Tarefas**: expressam conceituações sobre a resolução de problemas relacionados a uma atividade ou tarefa genérica, como

venda ou atendimento ao cliente, independente do domínio onde estas são aplicadas.

- **Ontologias de Aplicação**: são freqüentemente especializações das ontologias de domínio e tarefa específicas e relacionadas como venda de pacotes de viagem. Os conceitos contemplam, por exemplo, papéis associados aos atores do domínio ao desempenhar uma atividade. Devido ao seu nível de especialização, ontologias de aplicação são mais difíceis de serem reutilizadas.

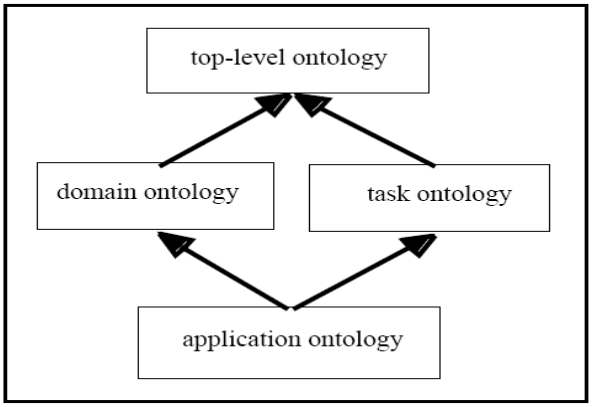


Figura 2.3 Relação de Generalização e Especialização entre Ontologias segundo (Guarino 1998).

2.2.2 Sistemas de Classificação e Indexação

Outra forma de classificação das ontologias diz respeito ao seu nível de formalidade para especificar os conceitos associados aos significados dos termos. Taxonomias, tesauros, glossários e outras estruturas de representação do conhecimento apresentadas à esquerda da figura 2.4 são considerados ontologias leves. Mas à medida que aumenta o grau de formalidade e expressividade das estruturas é possível reduzir a ambigüidade e aumentar a capacidade de raciocínio automatizado (Uschold e Gruninger 2004).

Glossário é uma lista fechada de termos que representam conceitos específicos, associados às suas definições, e que podem ser utilizados para classificação de objetos. Um anel de sinônimos permite que vários termos que representam um mesmo conceito sejam interligados por serem considerados equivalentes para propósitos de classificação e busca por recursos. Quando uma taxonomia é especificada, além dos termos que representam os conceitos que a compõem, são estabelecidas relações hierárquicas entre estes conceitos. A classificação de recursos de informação em relação a uma taxonomia permite o agrupamento de recursos de informação através de níveis hierárquicos superiores.

As normas ISO 2788 e ISO 5964 descrevem o padrão de construção de tesauros, sendo a primeira para tesauros de somente um idioma e a segunda para vários idiomas. Os tesauros estendem as taxonomias, pois além das relações hierárquicas entre os conceitos, também são acrescentadas outras relações entre os termos e outros construtos. Em geral um tesauro contém:

- BT/TG (*broader term* / termo geral): termo com significado mais geral. Um termo pode possuir mais de um BT.
- NT/TE (*narrower term* / termo específico): é o inverso de BT, representa um termo mais específico na escala hierárquica.
- SN (*scope note*): uma descrição associada ao termo para melhor elucidar o significado do conceito que o termo representa.
- USE (*use*): estabelece a relação de sinonímia entre os termos que se referem a um mesmo conceito e define que um termo é preferido em relação ao outro. A relação inversa é representada por UF (*used for*).
- TT (*top term*): termo que representa o ancestral mais geral da hierarquia.
- RT (*related term*): termo relacionado de outra forma que não por sinonímia ou hierarquia. Este elemento acrescenta semântica, porém não define que tipo de relacionamento existe entre os conceitos que os termos representam.

As relações hierárquicas (BT / NT) nos tesauros ainda podem ser refinadas em genéricas (entre classes) identificadas através das siglas BTG (*broader term generic*) e NTG (*narrower term generic*), instanciação como BTI (*broader term instance*) e NTI (*narrower term instance*) e parte-todo usando as abreviações BTP (*broader term partitive*) e NTP (*narrower term partitive*).

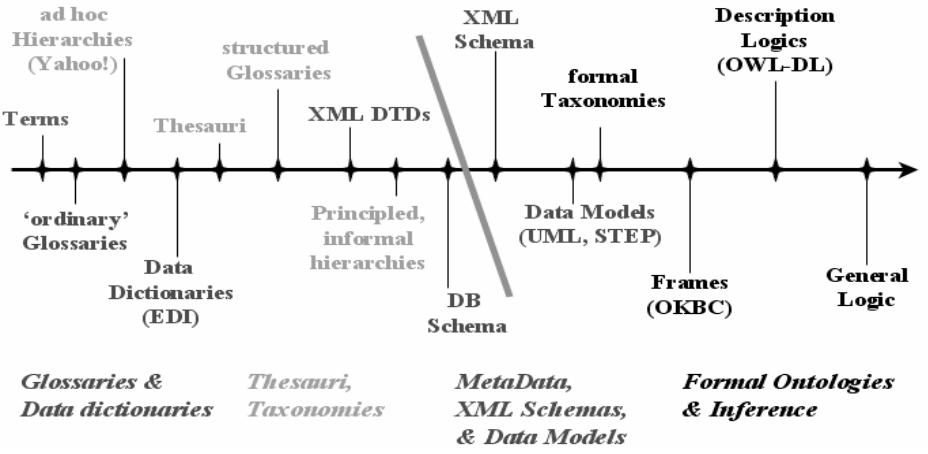


Figura 2.4 Estruturas de Representação do Conhecimento (Uschold e Gruninger 2004).

Glossários, anéis de sinônimos, taxonomias e tesauros também são conhecidos como linguagens de indexação e podem ser usados para evitar que os responsáveis por descrever o conteúdo dos recursos de informação escolham termos fora do contexto, no nível de especialização ou generalização inadequada e até formas diferentes do mesmo termo, como por exemplo, para o preenchimento do elemento “palavras- chave” do padrão de metadados Dublin Core.

Se comparadas com outros esquemas de classificação, as ontologias permitem a construção de modelos de domínio mais precisos e completos e a semântica dos dados recuperados se torna explícita (Sheth 1998), (Ziegler e Dittrich 2007). Além de relacionamentos semânticos padronizados, as ontologias também permitem a representação de relacionamentos específicos do domínio e de atributos dos conceitos (Breitman *et. al.* 2007).

Essas características as tornam recursos promissores para descrever o vocabulário de metadados, fornecendo uma estrutura semântica rica e formal para sua interpretação. Mas seu uso ainda requer esclarecimentos: Como esquemas de metadados farão uso de ontologias? Como as ontologias podem ser usadas para expressar necessidades de informação que demandam por localização de recursos de informação? (Sicilia 2006).

2.2.3 Similaridade Semântica

Ontologias são sistemas de sinais utilizados para comunicação e por isso também são avaliadas sobre o aspecto da semiótica. Semiótica é o estudo dos sinais e das formas pelas quais os sistemas de sinais são usados para transmitir significado. A semiótica pode ser subdivida em quatro aspectos: (1) sintaxe: que símbolos existem, (2) semântica: qual o significado desses símbolos, (3) pragmática: como os sinais são usados para propósitos específicos e (4) social: quem usa quais sinais (Maedche e Staab 2001).

Do ponto de vista da lingüística, que restringe a semiótica ao estudo dos signos lingüísticos, a semântica refere-se ao estudo do significado das palavras, das frases e do texto dentro de um contexto, envolve a explicitação do conhecimento intuitivo de uma língua, das regras que presidem à construção de predicados e dos mecanismos que garantem a sequência de enunciados no plano discursivo.

As palavras ou termos representam conceitos e estes por sua vez são abstrações de objetos do mundo real. Algumas palavras podem compartilhar o mesmo campo lexical, quando possuem o mesmo radical e que por isso tem um significado comum que advém deste radical ou o mesmo campo semântico, onde unidades lexicais com lexemas diferentes estão unidas pela proximidade de seu sentido. Além

disto, as relações semânticas entre as palavras, ou entre os conceitos que elas representam, podem ser hierárquicas, de inclusão, de equivalência ou de oposição (Lopes e Rio-Torto 2007).

Medidas quantitativas para calcular a similaridade entre termos foram propostas na literatura, algumas medem a similaridade léxica baseando-se na grafia da palavra, como por exemplo, a distância de edição (ou algoritmo de *Levenshtein*) e suas variações como *Hamming* (Navarro 2001). Outras medem a similaridade semântica usando corpus de documentos, como por exemplo, *Pointwise mutual information* (PMI-IR), ou estruturas de representação do conhecimento como ontologias (Mihalcea *et. al.* 2006).

A similaridade semântica baseada em estruturas de representação do conhecimento compara o significado, pois é baseada nos conceitos que os termos representam e lida com diferentes níveis de granularidade e abstrações. Estes métodos são classificados por Petrakis *et. al.* (2006) em quatro abordagens:

(1) contagem de arestas: a similaridade entre dois conceitos é calculada de acordo com o número de arestas do menor caminho que liga estes conceitos considerando a posição destes na taxonomia,

(2) conteúdo de informações em comum: mede a diferença entre dois termos através da probabilidade de ocorrência em um corpus, sendo que termos que denotam conceitos mais gerais que possuem muitos hipônimos apresentam menos conteúdo de informação do que termos referentes a conceitos mais específicos,

(3) baseado em características: as características em comum como atributos e relacionamentos tendem a aumentar a similaridade e

(4) híbrida que combina parcial ou integralmente as opções anteriores.

Essas abordagens podem ser aplicadas a uma única ontologia ou entre ontologias sendo a contagem de arestas e o cálculo do conteúdo de informações em comum mais indicados quando se trata de uma única ontologia.

Para permitir o reuso e expansão do conhecimento representado através de mais de uma ontologia existente é necessário identificar as similaridades e diferenças entre elas. Ontologias podem ser criadas em diferentes linguagens de representação, o que requer um processo de normalização antes de realizar alguma operação de casamento entre elas.

Mesmo quando estas são criadas usando uma linguagem comum podem existir divergências como um símbolo lingüístico representando conceitos diferentes (homonímia), uso de símbolos diferentes para representar um conceito (sinonímia), uso de convenções de modelagem diferentes para o mesmo recurso (conceito x

atributo), cobertura de domínio mais ampla ou mais restrita, nível de granularidade diferente (especialização e generalização ou agregação), etc... (Noy 2004).

O resultado da descoberta de similaridades permite a geração de uma nova ontologia através de combinação ou integração das ontologias originais. A combinação de ontologias gera uma nova ontologia a partir de ontologias similares ou sobrepostas sem manter uma definição clara da origem de cada recurso. Já a integração permite identificar a origem de cada recurso presente na ontologia resultante criada por adaptação, extensão ou especialização das ontologias originais que não precisam necessariamente pertencer ao mesmo assunto. Nos casos onde o objetivo não é a criação de uma nova ontologia e as ontologias originais são mantidas, as similaridades geram os alinhamentos ou mapeamentos entre elas (Choi *et. al.* 2006).

2.2.4 Ontologias Aplicadas à Integração de Informações

Ao analisar o impacto do uso de uma ontologia explícita sobre sistemas de informação, podem ser identificadas duas dimensões: uma dimensão temporal, que considera se a ontologia é utilizada em tempo de desenvolvimento ou em tempo de execução, e uma dimensão estrutural, relativa à forma particular como uma ontologia pode afetar os componentes de um sistema de informação (programas, fontes de informação e interfaces) (Guarino 1999). Ao utilizar uma ontologia em tempo de execução podem ser distinguidos dois comportamentos: ciente de ontologia (*ontology-aware*) e dirigido a ontologia (*ontology-driven*).

Ontologias podem ser usadas em tempo de desenvolvimento de SII como um modelo conceitual para a criação dos mapeamentos entre os esquemas conceituais heterogêneos e também em tempo de execução com um esquema mediado para suportar as consultas dos usuários e aplicações. A figura 2.5, extraída de (Wache 2001), apresenta uma taxonomia para classificação de SII de acordo com a abordagem de uso de ontologias.

A primeira abordagem (a) usa uma ontologia global e todas as fontes de dados são relacionadas a esta. A ontologia possui um vocabulário compartilhado para a especificação da semântica das fontes. Esta abordagem é aplicável em cenários onde é possível integrar todas as fontes de informação através da mesma conceitualização do domínio. A ontologia global pode ser gerada a partir da integração de outras ontologias.

Na abordagem de múltiplas ontologias (b), cada fonte de informação é descrita por sua ontologia local e por isso não é necessária a criação de uma ontologia global. Porém é necessário criar alinhamentos entre as ontologias específicas, pois a correspondência semântica de recursos das diferentes fontes ontológicas é uma

premissa para integração e a falta de um vocabulário comum torna a criação dos alinhamentos uma tarefa difícil.

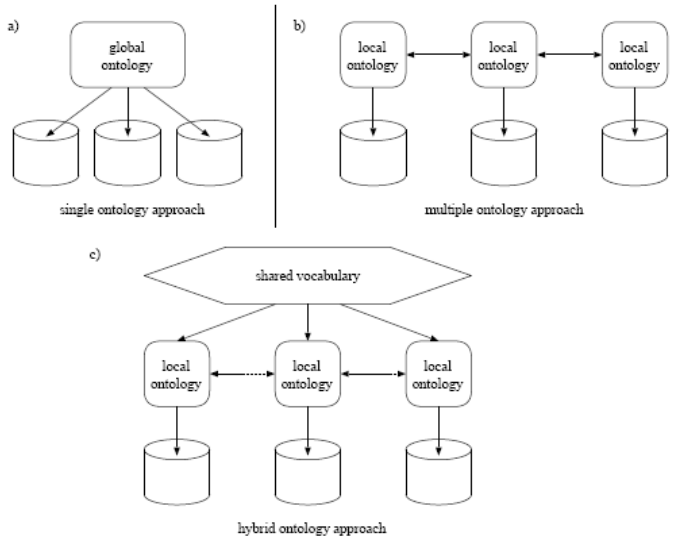


Figura 2.5 Abordagens de Uso de Ontologias para Integração de Dados (Wache 2001).

Na abordagem híbrida (c), cada fonte possui sua própria ontologia para deixar clara a semântica do seu conteúdo, porém estas ontologias são construídas levando em consideração um vocabulário comum que pode inclusive ser outra ontologia de mais alto nível. Nesta abordagem, a dificuldade está em reutilizar ontologias existentes.

2.3 Sistemas de Recuperação de informação

Sistemas de recuperação de informações (SRI) são projetados para encontrar objetos digitais que satisfazem uma necessidade de informação e estão armazenados em grandes coleções (Manning *et. al.* 2009). A busca é feita usando conteúdo não-estruturado, normalmente descrito em linguagem natural, e especificando palavras-chave que representam esta necessidade de informação e a ferramenta de busca retorna uma lista de objetos digitais. Estes objetos digitais são documentos, tabelas, gráficos, vídeos ou imagens, mas, em contextos mais específicos de acordo com o objetivo da coleção de objetos, podem corresponder a outros recursos como metadados de fontes de informações.

Um modelo muito utilizado em SRIs é o clássico. Neste modelo cada documento D_i é tratado como um conjunto de termos, conhecido como *bag of words*, sem considerar a ordem das palavras, associados a pesos onde $D_i = \{t_1 (w_1), t_2 (w_2),..., t_n (wn_n)\}$. O peso define a importância da palavra para descrever o objeto, quando o termo não aparece no conteúdo, o peso associado é zero.

O cálculo do peso do termo pode ser atribuído pelo número de ocorrências deste no conteúdo, esta medida é conhecida como *term frequency* (TF). Outra opção é a medida *term frequency–inverse document frequency* (TF-IDF) que corresponde à divisão da freqüência do termo (TF) pela *inverse document frequency* (IDF) que avalia a importância do termo na coleção. O cálculo da IDF é realizado obtendo o logaritmo do resultado da divisão do total de documentos pelo número de documentos contendo o termo. Desta forma a seletividade de uma palavra aumenta proporcionalmente com o número de vezes que a palavra ocorre no documento, mas diminui quanto maior for a sua freqüência na coleção. Esta forma de atribuição de peso requer que o cálculo seja refeito quando novos documentos forem acrescentados na coleção.

No modelo booleano, uma variação do modelo clássico, além das palavras-chave, é possível especificar operadores lógicos como E, OU, NÃO, mas o critério de seleção binário não permite a realização do casamento parcial entre estas palavras e os documentos e nem a definição de um critério de ordenação dos resultados.

No modelo vetorial, outra variação do modelo clássico, um peso não-binário é atribuído a cada termo e armazenado junto ao mesmo, em forma de um vetor. A medida de similaridade entre documentos d_i e d_j ou entre uma consulta q e um documento d, conhecida como similaridade do cosseno, é calculada através do cosseno do ângulo formado entre seus vetores em um espaço n-dimensional. Esta medida varia entre 0 e 1 e quanto mais próximo de 1, maior a similaridade entre eles, , além disto, permite casamento parcial entre consulta e documento e a ordenação do resultado pela pontuação em ordem decrescente.

Para geração dos termos de indexação, que representam o conteúdo da coleção de documentos de modo sumarizado (também são considerados metadados), algumas etapas de pré-processamento são necessárias. Na primeira etapa cada documento é separado em *tokens*, ou seja, em seqüências de caracteres agrupadas que representam uma unidade semântica de processamento, e são removidos os caracteres de pontuação.

Na etapa seguinte é realizada a remoção de *stopwords*, que são palavras que tem pouco valor para seleção de documentos. A lista de palavras pode ser determinada através da identificação de palavras com maior freqüência na coleção de documentos e também por uma lista padronizada que contenham preposições, conjunções, numerais. A normalização, onde as palavras são reduzidas a sua forma canônica, é a terceira etapa.

A última etapa é apoiada por algoritmos de *stemming* e/ou *lemmatization*. O primeiro é um processo heurístico que remove o fim das palavras, para converter plural em singular, e os afixos (sufixos e prefixos). Já o segundo usa um vocabulário

de apoio e faz análise morfológica das palavras com o objetivo de retornar a base ou a forma no dicionário de uma palavra, também conhecido como lexema ou semantema. O objetivo é reduzir as palavras ao radical, pois este é o elemento portador de significado, comum a um grupo de palavras da mesma família. Através destes algoritmos, que dependem da língua onde está sendo utilizado e da cobertura do dicionário, é possível identificar palavras semelhantes mesmo com grafias diferentes devido ao acréscimo de sufixos, prefixos, plural, conjugação de verbos.

No entanto, esses termos usados para indexação podem não corresponder exatamente aos utilizados para realizar a busca uma vez que estes são descritos em linguagem natural. As variações na expressão do mesmo conceito com diferentes termos (sinonímia) e múltiplos significados para o mesmo termo (polissemia) requer que a busca considere a similaridade no nível do significado, pois similaridade léxica isoladamente não identifica todas as possibilidades de aproximação.

2.3.1 Medidas de Desempenho

A relevância é uma característica central de SRI que distingue de Sistemas de Recuperação de Dados (Manning *et. al.* 2009). Um documento é relevante se atende a uma necessidade de informação e a identificação de relevância depende que um especialista indique quais documentos são relevantes entre todos os da coleção. O número de documentos relevantes recuperados determina o desempenho do SRI.

Para avaliar a qualidade do resultado recuperado são utilizadas duas medidas: precisão e cobertura. Precisão é calculada pela razão entre o total de documentos relevantes recuperados e o total de documentos recuperados e o seu objetivo é avaliar qual fração dos resultados atende a necessidade de informação. Cobertura é calculada como a razão entre o total de documentos relevantes recuperados e o total de documentos relevantes existentes no repositório de documentos e o seu objetivo é avaliar qual fração dos documentos que atendem a necessidade de informação é recuperada pela ferramenta de busca. O valor máximo para ambas as medidas é 1.

A vantagem da adoção das duas medidas é que em algumas circunstâncias uma é mais importante que a outra. Em buscas na web, os usuários preferem que os documentos relevantes sejam retornados nas primeiras posições, mas não se importam com a quantidade de documentos recuperados e nem se todos são relevantes, o que sugere que a precisão é mais importante neste caso. Em outras atividades, os usuários estão interessados que todos os documentos relevantes sejam recuperados, nestes casos o objetivo é maximizar a cobertura, mesmo que para isto tenham que analisar uma quantidade maior de documentos. A medida F estabelece

um *trade-off* entre precisão e cobertura por se tratar um uma média harmônica ponderada entre elas.

O resultado recuperado é ordenado de acordo com a medida de similaridade associada a cada documento em relação ao conjunto de palavras-chave que compõem a consulta. Na maior parte das abordagens este cálculo considera as medidas TF ou TF-IDF e a similaridade do cosseno entre a consulta *q* e cada documento *d*. Se o cálculo considerar o posicionamento das palavras no texto, a proximidade entre os termos dentro de um documento aumenta a pontuação do documento no resultado.

2.3.2 Abordagens para Expansão de Consultas

Técnicas de expansão de consultas e técnicas de redução de dimensão procuram diminuir as chances de consultas e documentos referenciarem o mesmo conceito utilizando diferentes termos (Manning *et. al.* 2009).

A redução de dimensão pode ser feita através de Análise da Semântica Latente (LSA) que é uma técnica estatística para extrair dos documentos uma estrutura latente de utilização das palavras, parcialmente oculta devido à sua variabilidade para descrever o mesmo tópico. Esta técnica pressupõe que a co-ocorrência de palavras reflete sua similaridade, faz uso de corpus de documentos e medidas de proximidade semântica entre termos e estende o modelo vetorial através de uma aproximação reduzida por decomposição de valores singulares (SVD) da matriz termo-documento que representa o corpus.

A expansão de consultas pode ser apoiada pela utilização de métodos locais, quando a consulta é expandida usando a lista inicial de documentos recuperada e refinamentos iterativos da consulta original, com ou sem apoio do usuário para seleção dos documentos relevantes, são realizados pelo sistema (Manning *et. al.* 2009). A abordagem de expansão de consulta baseada em modelos de conhecimento dependentes do corpus de documentos permite utilizar padrões de co-ocorrência de termos que compartilham o mesmo contexto para a geração da lista de palavras-chave expandida (Bhogal *et. al.* 2007). Outras abordagens usam métodos globais (Manning *et. al.* 2009) com apoio de dicionários de sinônimos, dicionários controlados com a forma canônica dos termos, navegação em taxonomias ou tesauros, que são modelos de conhecimento independentes do corpus de documentos (Bhogal *et. al.* 2007), para modificar a consulta original permitindo a combinação com outros termos semanticamente similares.

Sistemas de classificação acrescentam palavras ao conjunto de palavras original ou permitem a navegação pela coleção quando previamente classificada. Os

anéis de sinônimos, por exemplo, melhoram o resultado das consultas quando o problema da sinonímia ocorre na coleção. Taxonomias são usadas para guiar o usuário no refinamento das consultas através de conceitos mais gerais (*zoom out*) ou mais específicos (*zoom in*). Tesauros agregam as funcionalidades anteriores, acrescentam a opção de navegação por relacionamentos associativos entre os conceitos e aumentam o entendimento humano dos conceitos através de suas descrições (*scope notes*).

2.3.3 Busca Semântica

Algumas propostas para adaptação de ferramentas de busca utilizam ontologias e mecanismos de inferência para melhorar a precisão e cobertura do resultado e minimizar a interação do usuário (Mangold 2007)**.** Na busca semântica, o processo de recuperação de documentos explora o conhecimento do domínio e o uso de ontologias se justifica porque estas servem para compartilhar a mesma estrutura de informação entre pessoas e agentes de software, permitindo o reuso do conhecimento. As abordagens de busca semântica podem ser categorizadas conforme proposto por Mangold (2007), segundo as características apresentadas na figura 2.6 e descritas a seguir.

Na arquitetura *stand-alone* o sistema armazena os metadados dos documentos em uma estrutura semântica de indexação local que é usada para atender às solicitações de consulta. Na arquitetura *meta-search,* as consultas são distribuídas para outras ferramentas de busca subordinadas e o resultado é combinado em seguida antes de ser apresentado ao usuário.

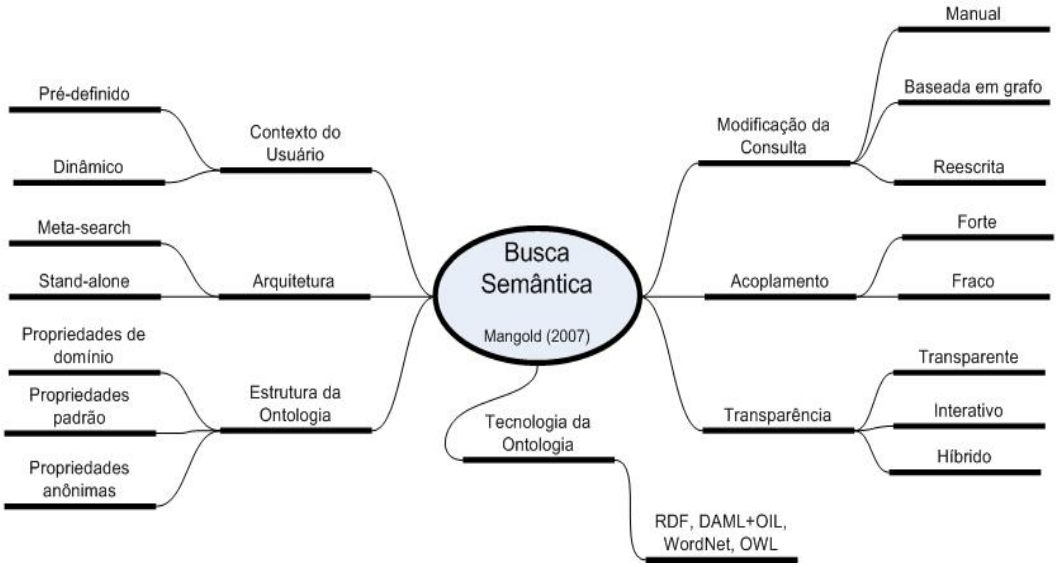


Figura 2.6 Critérios de Classificação de Ferramentas de Busca Semântica segundo (Mangold 2007)

O acoplamento entre a ontologia e os documentos pode ser forte ou fraco. No primeiro caso, os metadados de um documento se relacionam explicitamente a uma ontologia, pois o documento é representado como instância da ontologia e isto requer a anotação semântica dos documentos. As anotações semânticas são informações adicionais anexadas aos documentos que identificam ou definem o conceito em um modelo semântico que descreve parte do documento. A geração destas anotações é uma tarefa dispendiosa, mas permite métodos de recuperação melhores em termos de precisão e cobertura dos resultados. O acoplamento forte permite medir o quanto dois documentos são semanticamente similares através da distância conceitual, considerando o menor caminho entre os conceitos que estes documentos estão associados como instâncias.

O contexto do usuário pode ser extraído do histórico de interações do mesmo, o que caracteriza um sistema dinâmico, mas pode ser do tipo pré-definido através de uma lista de categorias de perguntas, onde há menos flexibilidade e por isso é mais indicado para domínios específicos.

A transparência acontece quando não são solicitadas informações adicionais pelo sistema ao usuário de modo a esclarecer o contexto da busca. Quanto mais transparente, menos o usuário pode interferir nas decisões do sistema.

A modificação da consulta pode ser:

(1) manual, quando a consulta retorna documentos e um trecho da ontologia associado e o usuário refina a consulta através da própria navegação pela ontologia;

(2) reescrita por reformulação da lista de palavras-chave através de:

argumentação, onde novos termos, relacionados com os conceitos na ontologia, são derivados para aumentar a abrangência da consulta,

substituição dos termos da consulta pelos termos que identificam os conceitos da ontologia que são sinônimos, hiperonímia e hiponímia e

remoção de termos, que possuem baixa seletividade na coleção e

(3) baseada em grafo, quando existe acoplamento forte entre a ontologia e os documentos e o algoritmo percorre o grafo para recuperar documentos que são instâncias de conceitos semanticamente relacionados.

As propriedades de objetos presentes na estrutura da ontologia podem ser anônimas, padronizadas e dependentes de domínio. As propriedades anônimas só indicam que os conceitos compartilham o mesmo contexto sendo as propriedades padronizadas e dependentes de domínio mais indicadas para buscas semânticas, pois estas explicitam o tipo de relação que existe entre os conceitos.

2.3.4 Busca e Exploração

As necessidades de informação que motivam a busca são as mais variadas, por isso a atividade de busca por informação pode ser separada em três grupos: pesquisa (*lookup*), aprendizado (*learn*) e investigação (*investigate*). Os tipos aprendizado e investigação são considerados como busca exploratória. Mas estas não são mutuamente exclusivas, ou seja, o usuário pode realizá-las de modo concomitante ou alternando entre elas (Marchionini 2006).

Pesquisa tem por objetivo a recuperação de fatos e é adequada para situações onde a busca analítica começa com consultas criteriosamente especificadas, recuperam resultados precisos e requerem pouco esforço de avaliação e comparação entre os itens recuperados. Os resultados são objetos discretos e bem estruturados como nomes, números, frases curtas ou objetos digitais específicos do tipo texto ou outras mídias.

A busca motivada pelo aprendizado aumentou o seu potencial à medida que novos recursos são disponibilizados. Esta atividade envolve algumas iterações e os resultados são submetidos a processamento cognitivo e interpretação. A maior parte do tempo com este tipo de busca é aplicado na avaliação e comparação dos itens e na reformulação das consultas para descobrir as fronteiras do significado dos conceitos. O aprendizado é aplicável para combinar estratégias de análise e navegação nos resultados com pesquisa, pois este último permite recuperar os conceitos vizinhos corretamente.

Investigação é um processo de busca longo que envolve muitas iterações e os resultados recuperados são avaliados de modo mais crítico antes de serem incorporados no corpo de conhecimento. Este tipo de busca é aplicável para suportar atividades de planejamento ou para transformar dados existentes em novos dados e conhecimento novo. Neste caso, o objetivo é maximizar o número de objetos relevantes que são recuperados, ou seja, aumentar a cobertura ao invés da precisão. As máquinas de busca atuais não são adequadas para investigação pois priorizam a precisão nas primeiras páginas de resultado, o que explica a causa de novos serviços especializados de busca estarem sendo propostos.

A busca para aprendizado e para investigação depende fortemente da participação humana neste processo contínuo de recuperação e análise dos resultados. Neste cenário, recuperar informações requer mais que tratar o problema de combinar consultas e documentos e retornar o resultado ordenado de acordo com uma pontuação relativa a esta similaridade. A recuperação de informações deve considerar a interação do agente humano, suas necessidades de informação e seu

perfil assim como a capacidade do repositório de informações para atender a estas necessidades e a sua dinâmica de agregar novas fontes de informações além da evolução destes ao longo do tempo.

2.4 Considerações Finais

O nível de maturidade da Arquitetura de Informações de uma Organização permite a avaliar o uso do acervo de informações e o valor que estas práticas agregam ao negócio. A maturidade aumenta a medida que: (1) um programa de Governança de Informações é implantado, (2) processos de integração de informação e de avaliação e melhoria da qualidade de dados são definidos, executados e monitorados e (3) existe uma estratégia para gerenciamento de metadados por todo o seu ciclo de vida (Godinez *et. al.* 2010).

Uma vez que a maioria dos padrões de esquemas de metadados é composta por elementos contendo descrições em linguagem natural, o que permite o uso e interpretação humana do seu conteúdo, também é possível aplicar buscas por palavras-chave em repositórios de metadados. Para garantir qualidade de conteúdo na descrição dos elementos, devem ser utilizados vocabulários controlados (NISO 2004). Os resultados recuperados estarão relacionados com o assunto de interesse quando os termos usados correspondem àqueles utilizados na busca.

A busca por fontes de informação visa obter um panorama das fontes disponíveis (Bernstein e Haas 2008) que estão registradas no repositório de metadados e serve como ponto de partida para exploração de dados. Mas, assim como na busca por documentos, esta busca pode apresentar baixa precisão (número de fontes de informações relevantes recuperadas / número de fontes de informações recuperadas) e cobertura (número de fontes de informações relevantes recuperadas / número de fontes de informações relevantes existentes na organização). Isto pode acontecer em função de:

(1) divergência entre os termos da consulta e os utilizados no conteúdo dos elementos descritivos do esquema de metadados;

(2) conflitos, de nomenclatura e estrutura, devido aos diferentes contextos em que as fontes de informação foram modeladas;

(3) perda semântica dos processos de modelagem, desenvolvimento e catalogação das fontes de informação e

(4) nem todas as fontes de informação relevantes existentes na organização estão registradas no repositório de metadados.

O objetivo desta pesquisa acadêmica é propor uma arquitetura que torne a busca por fontes de informação mais eficiente, facilitando o processo de integração e o reuso das mesmas dentro da organização. Para isto, é necessário identificar a semântica da consulta, direcioná-la às fontes relevantes desta coleção, tornar explícita a semântica destas fontes e a similaridade semântica entre elas.

No próximo capítulo serão descritos os componentes de uma arquitetura lógica que explora o conhecimento do domínio, modelado através de uma ontologia, e utiliza busca semântica no repositório de metadados corporativo. Esta arquitetura permite minimizar os problemas decorrentes de divergência terminológica, conflitos das fontes de informação e perda semântica durante a busca por fontes de informação em repositórios de metadados.

3. Proposta de Solução

A atividade de busca por fontes de fontes de informação é a etapa posterior ao levantamento dos requisitos que especificam uma necessidade de informação. O resultado desta busca pode determinar se a fonte de informação que atende a esta necessidade já existe na organização e pode ser reutilizada, ou se deve ser desenvolvida ou adquirida externamente. Mas se esta demanda envolve o acesso, combinação e apresentação de informações de mais de uma fonte então é iniciado um processo de integração de informações. Em todos esses cenários, a atividade de busca depende do conhecimento do domínio e sua eficiência está relacionada com a existência, a completude e a capacidade de exploração do repositório de metadados onde as fontes de informação são registradas e semanticamente descritas.

Neste capítulo serão descritos os componentes de uma arquitetura, dirigida à ontologia, que permite identificar fontes de informação semanticamente similares ou relacionadas a serem integradas. Nesta arquitetura, ontologias de domínio, representam o conhecimento do domínio e são utilizadas para realizar busca semântica no repositório de metadados corporativo.

3.1 Visão Geral da Arquitetura Proposta

Uma arquitetura define a maneira como os componentes de computadores ou sistemas se organizam e se integram (Merriam-Webster 2005). De acordo com a norma IEEE 1471-2000, arquitetura é definida como a organização fundamental de um sistema, incorporada em seus componentes, na relação entre eles e o seu ambiente, e nos princípios que orientam o seu projeto e evolução.

De modo mais específico, arquitetura de software foi definida por Bass *et. al.* (1998) como a estrutura de sistemas que abrangem componentes de software, as propriedades destes que são externamente visíveis e a relação que é estabelecida entre eles. O objetivo da arquitetura de software (Garlan 2000) é tornar explícitas as decisões de projeto de dados e sistemas em um nível mais alto de abstração, facilitando a compreensão do seu funcionamento, da composição dos elementos, de

como a interação entre estes componentes ocorre e na descrição das funcionalidades dos mesmos. Uma arquitetura caracteriza uma família de sistemas e não apenas um sistema, além de auxiliar no gerenciamento da complexidade e justificar as escolhas arquiteturais específicas.

A arquitetura proposta é apresentada através do diagrama de componentes da UML na figura 3.1, cujos componentes são descritos a seguir:

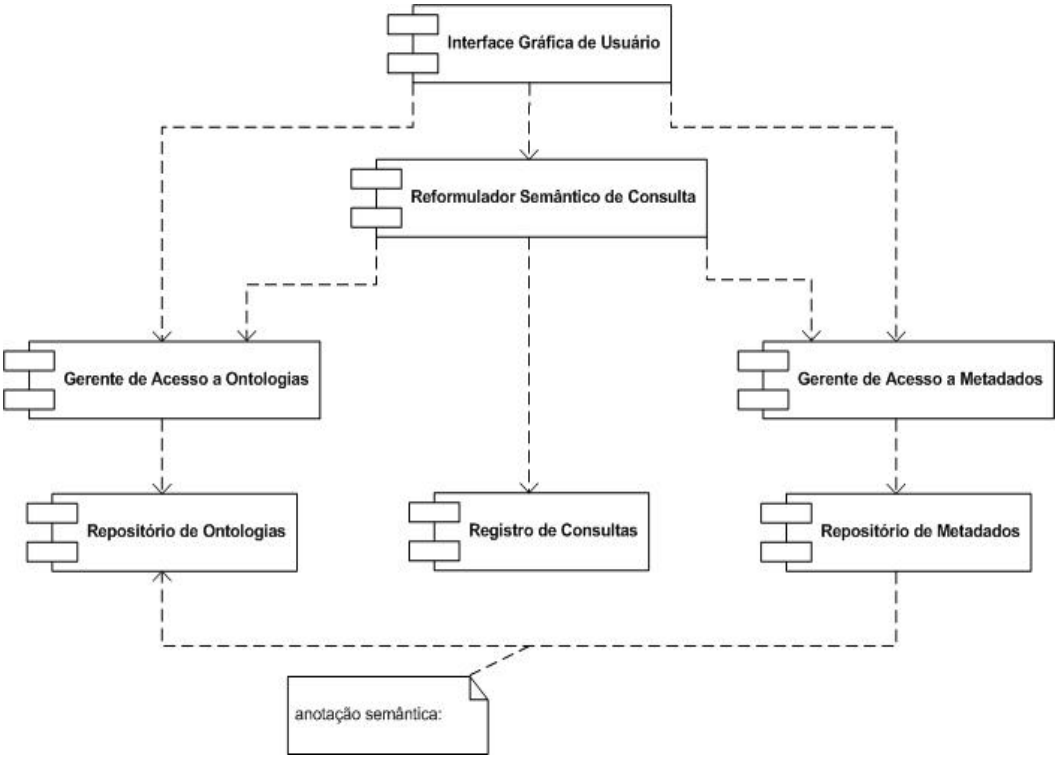


Figura 3.1 Componentes da Arquitetura Dirigida a Ontologia

1) **Interface Gráfica de Usuário**: este componente é responsável pela (i) visualização das ontologias de domínio através da interação com o componente Gerente de Acesso a Ontologias, (ii) entrada de dados e exibição do conteúdo dos metadados descritivos, administrativos e estruturados e das anotações semânticas das fontes de informação registradas, quando existirem, através da interação com o componente Gerente de Acesso a Metadados e (iii) por receber a lista inicial de palavras-chave e critérios de qualidade das fontes para a realização da busca por fontes de informação através da interação com o Reformulador Semântico de Consulta;

2) **Gerente de Acesso a Metadados:** todo acesso ao Repositório de Metadados para catalogação e anotação semântica das fontes de informação e a recuperação dos metadados destas fontes é realizado por este componente;

3) **Gerente de Acesso a Ontologias:** todo acesso ao Repositório de Ontologias para permitir a visualização das ontologias de domínio e a seleção de recursos para anotação semântica das fontes de informação e reformulação semântica de consultas é realizado por este componente;

4) **Repositório de Ontologias**: este repositório armazena as ontologias utilizadas para anotação semântica e busca por fontes de informação;

5) **Repositório de Metadados:** este repositório armazena os metadados descritivos, administrativos e estruturados de todas as fontes de informação registradas e as anotações semânticas destas fontes, quando existirem;

6) **Reformulador Semântico de Consulta:** a lista inicial de palavras-chave informada para busca por fontes de informação é transformada por este componente em duas consultas que recuperam os metadados das fontes de informação através de anotações semânticas e do conteúdo dos elementos descritivos. Este componente interage com o Gerente de Acesso a Ontologias para obter os URIrefs e rótulos dos recursos das ontologias de domínio a partir da lista inicial de palavras-chave e com o Gerente de Acesso a Metadados para recuperar os metadados das fontes de informação. Este componente também realiza a ordenação do resultado considerando as características de qualidade das fontes de informação que atendem a necessidade de informação.

7) **Registro de consultas:** todas as consultas realizadas através da funcionalidade de busca devem ser registradas em um *log*. Este *log* armazena a lista inicial de palavras-chave, os recursos das ontologias de domínio selecionados pelos usuários e o julgamento de relevância das fontes de informação em cada consulta. Este *log* fornece insumos para os processos de Engenharia de Ontologias, Governança de Informações e de Anotação Semântica de fontes de informação catalogadas.

O **Repositório de Ontologias** e o **Repositório de Metadados** foram modelados como componentes especializados, mesmo que ontologias sejam consideradas como metadados, uma vez que os repositórios possuem objetivos distintos e sistemas de registros de metadados possuem gerência somente sobre o **Repositório de Metadados**. Porém a ligação entre estes, através do ponteiro URIref dos recursos das ontologias de domínio associado às fontes de informação (**Anotação Semântica**), é uma característica essencial da arquitetura. Esta ligação permite que o conhecimento do domínio seja usado para aumentar a eficiência da busca por fontes de informação e para tornar explícita a semântica das fontes de informação registradas. A **Anotação Semântica** de uma nova fonte de informação, realizada em

conjunto com a sua catalogação no **Repositório de Metadados** pelo responsável pela mesma, evita a perda semântica no processo de catalogação uma vez que o responsável possui conhecimento sobre os conceitos, atributos e instâncias dos dados que estão armazenados nesta fonte. Mesmo assim, o processo de anotação semântica ainda deve ser conduzido para aumentar a precisão semântica de fontes de informação catalogadas à medida que as ontologias de domínio armazenadas no **Repositório de Ontologias** evoluem com a adição de novos recursos e geração de alinhamentos através de processos de Engenharia de Ontologias. Os recursos selecionados das ontologias de domínio e o julgamento de relevância das fontes de informação, fornecidos pelos usuários que realizam a busca e armazenados no **Registro de Consultas** da arquitetura, podem ser usado como insumos para o processo de anotação semântica desde que analisados pelos responsáveis pelas mesmas, seguindo uma abordagem semi-automática (Reeve e Han 2005).

A arquitetura proposta é caracterizada como dirigida a ontologia (Guarino 1998) uma vez que ontologias de domínio são utilizadas em tempo de execução pelas funcionalidades de busca e de anotação semântica das fontes de informação. A dinâmica da busca por fontes de informação é explicada a seguir e nos itens 3.3.1 até 3.3.5 são apresentados com mais detalhes cada um dos componentes da arquitetura proposta.

3.2 Busca por Fontes de Informação

O Repositório de Metadados é mantido por um sistema de informação de registro de metadados. A arquitetura apresentada propõe uma extensão da funcionalidade de busca por fontes de informação deste sistema. O processo de busca, ilustrado na figura 3.2, está dividido em duas etapas: (1) Busca por recursos (conceitos, instâncias, atributos e relacionamentos) nas ontologias de domínio e (2) Busca por fontes de informação no repositório de metadados. Na primeira etapa, o objetivo é identificar os conceitos, relacionamentos, atributos e instâncias que melhor representam a necessidade de informação do usuário. Já na segunda etapa, o objetivo é identificar dentre as fontes de informação que estão registradas aquelas que melhor atendem esta necessidade de informação.

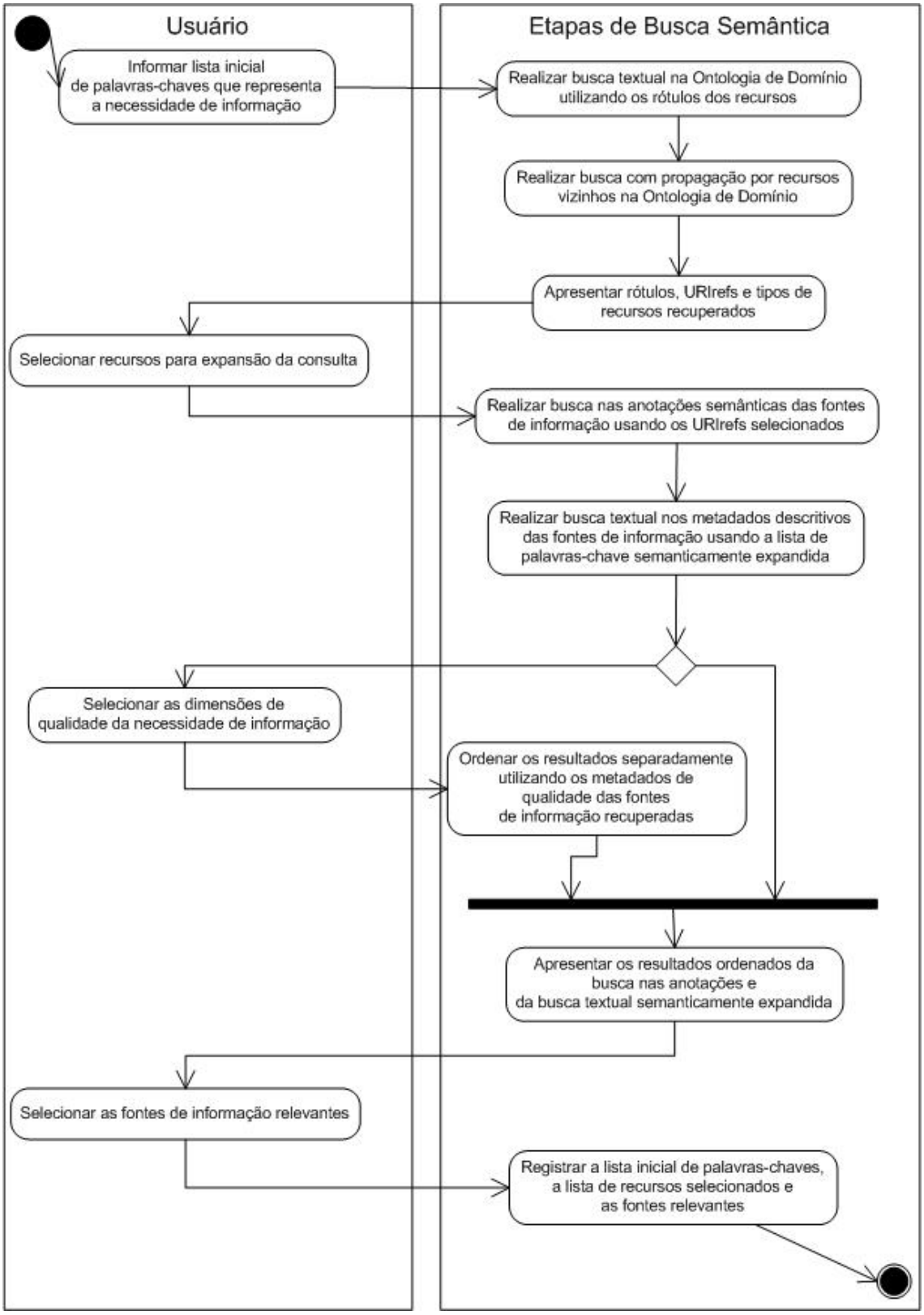


Figura 3.2 Sequência de Passos para Busca por Fontes de Informação

Através da Interface Gráfica de Usuário, o usuário especifica uma lista inicial de palavras-chave que representa a sua necessidade de informação. Esta lista é usada pelo Reformulador Semântico de Consultas para recuperar fragmentos das ontologias

de domínio, estes são subconjuntos dos recursos recuperados das ontologias de domínio a partir dos termos utilizados na consulta e de regras de propagação. Este componente interage com o Gerente de Acesso a Ontologias para obter rótulos e URIrefs de recursos das ontologias que foram selecionadas para a busca e catalogação das fontes e que estão armazenadas no Repositório de Ontologias.

Estes fragmentos são apresentados ao usuário e ele deve selecionar os URIrefs a serem utilizados e os rótulos a serem acrescentados nas consultas. Esta segunda interação do usuário é necessária para esclarecer o contexto da busca, pois a lista inicial de palavras-chave não é o suficiente para tornar explícitos os conceitos associados com a intenção de busca do usuário. O Reformulador Semântico de Consultas envia a lista de URIrefs ao Gerente de Acesso a Metadados para recuperar as fontes de informação do Repositório de Metadados através das Anotações Semânticas. A seguir o Reformulador Semântico de Consultas envia a lista de palavras-chave semanticamente expandida ao Gerente de Acesso a Metadados para recuperar os metadados das fontes de informação do Repositório de Metadados através do conteúdo dos elementos descritivos selecionados.

Os metadados recuperados são ordenados pelo Reformulador Semântico de Consultas de acordo com os critérios de qualidade selecionados pelo usuário e apresentados e analisados pelo usuário, este seleciona as fontes de informação que julgar relevantes, ou seja, aquelas que mais contribuem para atender a sua necessidade de informação. A lista inicial de palavras-chave, os recursos da ontologia selecionados pelo usuário e o julgamento de relevância são armazenados no Registro de Consultas. Se o usuário identificar que as fontes de informações recuperadas ainda não são suficientes, uma nova consulta deve ser realizada.

3.3 Detalhamento dos Componentes da Arquitetura Proposta

Cada componente da arquitetura lógica proposta é descrito nos itens a seguir para permitir o entendimento de suas funcionalidades e da interdependência entre estes componentes.

3.3.1 Interface Gráfica de Usuário

Através da interface gráfica é possível explorar o conhecimento do domínio representado pelas ontologias armazenadas no Repositório de Ontologias. A visualização de seus conceitos, hierarquia de conceitos, relacionamentos específicos de domínio, atributos e instâncias de conceitos, comentários e rótulos dos recursos permite que o usuário humano adquira conhecimento sobre o domínio ao interagir com

a interface. Porém, a visualização de ontologias não é uma tarefa fácil, uma vez que se trata de um artefato que representa muito mais que uma simples hierarquia de conceitos. A arquitetura não determina a forma de visualização que deve ser utilizada, funcionalidades como operações de navegação, aumento e redução do zoom, mudança de foco, rotação e filtro são desejáveis mas a operação mais importante para a arquitetura é busca textual nos rótulos dos recursos.

Katifori *et. al.* (2007) apresentam uma visão geral das técnicas de visualização e concluem que a escolha da abordagem apropriada deve levar em consideração tanto os requisitos funcionais, como a capacidade de navegação, quanto os não funcionais, como o tamanho das ontologias, além das tarefas específicas realizadas pela aplicação. Estas abordagens e as ferramentas que a implementam vêm sendo usadas no contexto de ferramentas de gerenciamento de ontologias, como por exemplo os diversos *plugins* do *Protégé*, mas também para auxiliar na recuperação de informações em aplicações de busca que utilizam ontologias. De acordo com os autores desse *survey*, a forma de visualização de ontologias pode ser dividida em seis categorias: lista indentada, nós com ligações (grafos) e hierarquia (árvores de taxonomia), capacidade de manipular o foco (*zooming*), subdivisão de espaços (*treemaps, information slices*), foco com contexto e distorção e paisagens em 3D.

O tipo de visualização mais utilizado é a **lista indentada**, disponível nas ferramentas *Protégé* [1], *OntoStudio* [2] e *Kaon* [3] . A visualização é realizada através de uma estrutura semelhante ao *Microsoft Windows Explorer* onde a taxonomia da ontologia (relações de hierarquia entre as classes) é representada em uma árvore que permite expandir ou contrair os nós. A maior vantagem deste tipo de visualização é a sua simplicidade de implementação e representação, além da familiaridade dos usuários, uma vez que este mesmo conceito é aplicado em vários outros softwares. Mas um problema desta abordagem é que ela representa uma árvore e não um grafo e com isso as relações não hierárquicas entre conceitos não são visualizadas nesta estrutura, além disto, nos casos de múltipla relação hierárquica esta relação acaba não ficando muito clara.

A segunda forma de visualização mais usada representa as ontologias como **um grafo** e também preserva o conceito visual de **uma árvore para a taxonomia**. *OntoViz* [4] é um exemplo de ferramenta que utiliza este tipo de visualização, trata-se de

[1] http://protege.stanford.edu/

[2] http://www.ontoprise.de/en/products/ontostudio/

[3] http://kaon.semanticweb.org/

[4] http://protegewiki.stanford.edu/wiki/OntoViz

um *plugin* do *Protégé* que utiliza a biblioteca *GraphViz* [5] para desenhar grafos em 2D. Cada classe é representada por um nó com seu nome, suas propriedades e as relações hierárquicas e papéis entre classes representam as ligações. O *RDF-Gravity* [6] é outro exemplo para visualização como um grafo mas não se trata de um plugin e sim uma ferramenta que permite a busca textual e filtros de visualização de recursos.

A capacidade de **manipular o foco** é encontrada em ferramentas como *Jambalya* [7] (outro *plugin* do *Protégé*) que apresentam os nós aninhados de acordo com a hierarquia dos conceitos e permitem que um nó seja selecionado por vez para aumentar o foco sobre este.

A **subdivisão de espaços** (*treemaps, information slices*) utiliza todo o espaço disponível para visualização e divide entre os nós de primeiro nível em seções com tamanhos variados. O tamanho do espaço de cada nó depende do número de nós que este contém e cada seção é subdividida sucessivamente entre os nós dos níveis seguintes.

A opção de visualização do foco com **contexto e distorção** é oferecida através de uma árvore hiperbólica. Esta abordagem fornece uma visão do contexto trazendo o foco para o recurso que está sendo analisado, que ocupa uma posição central, e distorcendo os demais através da redução do tamanho dos recursos próximos até os mais distantes atingirem um tamanho que não seja mais visível. *OntoRama* [8] é um exemplo de uma ferramenta que exibe uma base de conhecimento em RDF no formato de uma árvore hiperbólica.

A visualização na forma de **paisagens em 3D**, usada para visualização de documentos em sistemas de arquivos, não foi encontrada em nenhuma ferramenta de visualização de ontologias.

A interface gráfica também suporta as atividades de catalogação das fontes de informação, através do cadastramento de seus metadados descritivos, administrativos e estruturados e a associação das anotações semânticas, e de busca por fontes de informação. A interface de busca recebe a lista inicial de palavras-chave e os critérios de qualidade selecionados pelo usuário, exibe as sugestões de expansão baseadas em fragmentos recuperados da ontologia de domínio e apresenta as fontes de informação resultantes ordenadas segundo as características de qualidade das mesmas.

---

[5] http://www.graphviz.org/

[6] http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html

[7] http://protegewiki.stanford.edu/wiki/Jambalaya

[8] http://sourceforge.net/projects/ontorama/

A utilização de ontologias de domínio na busca e a visualização de seus fragmentos, através da interface gráfica, permitem que esta exerça o papel de um meio de comunicação entre o usuário e o sistema. De acordo com Guarino (1998), neste cenário o usuário está consciente da ontologia e a usa como parte do sistema de informação para melhor formular consultas no nível apropriado de detalhamento e entender o vocabulário usado. Desta forma, a ontologia torna-se uma ferramenta de aprendizado e disseminação do conhecimento do domínio.

3.3.2 Repositório de Ontologias e Gerente de Acesso a Ontologias

O Repositório de Ontologias pode conter todos os tipos de ontologias, mas somente as ontologias de domínio são usadas para anotação semântica e busca por fontes de informação dentro da arquitetura. A justificativa para esta escolha é que este tipo de ontologia modela classes e propriedades específicas de domínio e os rótulos de seus recursos são parte do vocabulário dos produtores e consumidores de informação dentro da organização. Ontologias linguísticas independentes de domínio como a *WordNet* foram evitadas para a busca e anotação devido a sua ampla cobertura (Bhogal *et. al.* 2007), além disso, estas ontologias não são apropriadas para acomodar jargões, neologismos e acrônimos que são constantemente agregados ao vocabulário dos usuários de uma comunidade ou organização.

Para permitir anotação e busca semântica nesta arquitetura, são utilizadas as ontologias de domínio representadas através da linguagem *OWL* (*Web Ontology Language*). A OWL foi selecionada porque, além de se tratar de uma recomendação da W3C para representação de ontologias na Web, esta linguagem permite a representação de classes (conceitos), propriedades de objetos e de dados (relacionamentos e atributos), axiomas (proposições lógicas) com expressividade além de conter construtos específicos para representar o papel dos conceitos nas relações (*rdfs:domain* / *rdfs:range*), as relações taxonômicas (*rdfs:subClassOf*) e restrições de valores que podem ser associados aos atributos de um objeto.

A arquitetura requer que os recursos da ontologia possuam pelo menos um rótulo em linguagem natural (*rdfs:label*) para que seja possível realizar o casamento das palavras-chave informadas com os recursos. Apesar de não contribuírem para a interpretação lógica da ontologia pelos motores de raciocínio, os rótulos apresentados permitem que o usuário humano esclareça o contexto da sua intenção de busca ao atribuir significado aos termos informados. Os comentários (*rdfs:comment*) são opcionais mas também podem contribuir para o entendimento dos conceitos e relacionamentos presentes na ontologia.

Além dos recursos recuperados através do casamento, parcial ou total, dos rótulos com os termos da lista inicial de palavras-chave, os recursos vizinhos também são recuperados ao expandir a busca na ontologia. Esta expansão é guiada por regras de propagação que exploram as relações padrão entre conceitos (como especialização e generalização, classes irmãs, composição ou agregação), as relações dependentes de domínio entre os conceitos, instanciação e atributos dos conceitos assim como a equivalência entre recursos de ontologias distintas.

As ontologias de domínio podem ter sido integradas ou combinadas para a criação de uma nova ontologia e alinhadas ou mapeadas para descobrir as relações de similaridades entre elas (Maedche e Staab 2002). Se ontologias de domínio forem submetidas à processos de combinação e as ontologias originais não forem mantidas no repositório, isto irá requerer que as fontes de informação sejam submetidas ao processo de anotação semântica novamente pois perderiam a ligação com o ponteiro dos recursos (URIrefs) das ontologias originais. A integração de ontologias que são similares ou sobrepostas mantém o ponteiro dos recursos (URIrefs) das ontologias originais e por isso não requer que as fontes sejam anotadas novamente. No caso do alinhamento, como representado na figura 3.3, a arquitetura utiliza o resultado deste processo ao explorar os construtos da *OWL* que representam a equivalência entre recursos (*owl:equivalentClass* / *owl:equivalentProperty*) através de regras de propagação. O resultado do mapeamento de ontologias não pode ser utilizado na arquitetura pois não são geradas ligações formais entre as ontologias que possam ser exploradas pelo Reformulador Semântico de Consultas.

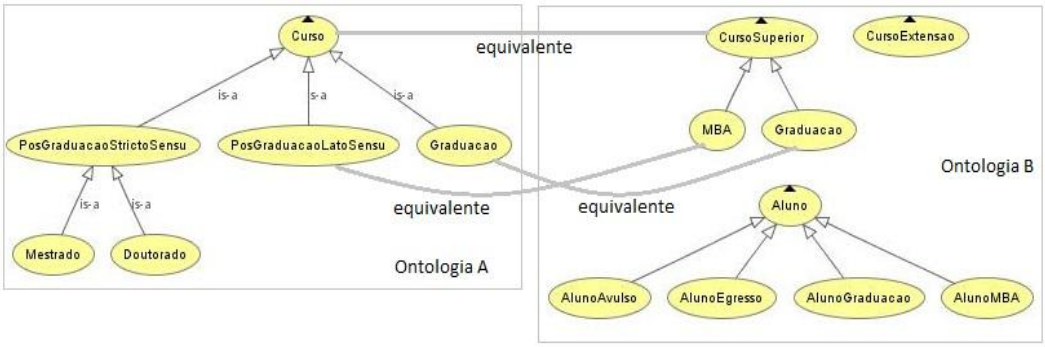


Figura 3.3 Alinhamentos que Representam Similaridade entre Conceitos

Uma vez que o(s) domínio(s) das fontes registradas no Repositório de Metadados é (são) conhecido(s) *a priori*, é possível selecionar um subconjunto de ontologias de domínio do Repositório de Ontologias e configurar o Gerente de Acesso a Ontologias de modo a utilizar somente este subconjunto a ser invocado pelo Reformulador Semântico de Consultas.

3.3.3 Repositório de Metadados e Gerente de Acesso a Metadados

O Repositório de Metadados é considerado a fonte ideal para realização da busca por fontes de informação, pois neste devem estar descritas todas as fontes de informações, internas e externas à organização, que são necessárias para suportar os processos de negócio. Dentro de uma iniciativa de definição da Arquitetura de Informações Empresarial, o gerenciamento de metadados através do repositório de metadados, se torna essencial para oferecer suporte às atividades de integração, gerenciamento, evolução e documentação de sistemas de informação.

Devido à heterogeneidade das fontes de informação existentes na organização, este repositório deve ser flexível a ponto de permitir que qualquer tipo de fonte possa ser registrada. Este requisito de flexibilidade vem motivando propostas de repositórios de metadados que fazem uso intensivo da linguagem XML e tecnologias associadas. Estas tecnologias facilitam o reuso e a configuração do software de gerenciamento de metadados. Ferreira e Moura-Pires (2007) apresentam uma solução para armazenamento, validação, consulta e transformação de metadados baseada em XML e serviços Web. O XML Schema é utilizado para definição e validação do esquema de metadados, a linguagem XSLT para transformação de documentos XML em outros formatos XML ou HTML, permitindo intercâmbio e apresentação, e a linguagem XQuery para consulta a estes documentos.

A arquitetura proposta não determina um esquema de metadados padrão, mas a utilização da arquitetura para estender um sistema de registro de metadados requer a identificação dos elementos que permitem a busca e a ordenação pelos critérios de qualidades. O gestor do repositório de metadados deve analisar a função dos elementos de cada esquema de metadados registrado e identificar aqueles que têm por objetivo a descrição das fontes de informação para fins de localização e os que descrevem as características de qualidade das fontes.

A busca textual em metadados utiliza os elementos descritivos, uma vez que estes revelam parte da semântica das fontes, como a intenção de uso, sua descrição ou resumo do seu conteúdo (Sheth 1998), fornecendo o contexto para entendimento dos dados através do tempo. Um aspecto crucial na criação de elementos descritivos diz respeito ao vocabulário e muitos esquemas de metadados utilizam sistemas terminológicos para descrição. Ontologias também podem ser usadas na criação destes elementos, pois contêm termos de um vocabulário controlado para representar conceitos e suas relações (Kashyap *et. al.* 2008) e estes mesmos rótulos são usados na arquitetura proposta para a expansão semântica da lista de palavras-chave.

No entanto, para garantir maior precisão na descrição das fontes de informação e conseqüentemente nos resultados da busca, a ontologia deve ser usada como um modelo de referência para anotação, ligando as fontes de informação e seus metadados aos recursos da ontologia através do ponteiro URIref. Tais Anotações Semânticas fazem referência a recursos das ontologias existentes no Repositório de Ontologias para descrever o conteúdo das fontes de informação catalogadas no Repositório de Metadados. Dessa forma, as ontologias de domínio são usadas para organizar e classificar o conteúdo do Repositório de Metadados em um nível mais alto de abstração.

O esquema de metadados deve contemplar um elemento com o propósito de armazenamento das Anotações Semânticas no Repositório de Metadados para permitir a busca utilizando anotações. Este elemento deve permitir múltiplas ocorrências e o seu conteúdo corresponde ao ponteiro para um recurso de uma ontologia. Mas este elemento não pode ser obrigatório para permitir que os metadados de fontes de informações legadas, que não contêm anotação, possam ser armazenados junto com os metadados das novas fontes de informação que venham a ser descritas e anotadas. Avaliar ou propor métodos de anotação semântica não faz parte do objetivo principal da arquitetura proposta, mas é importante considerar que a anotação semântica permite melhor precisão nos resultados da busca.

Os aspectos de qualidade de dados importantes para uma determinada necessidade de informação devem ser especificados pelo usuário. Estes correspondem a um subconjunto das medidas qualitativas e quantitativas (Batini e Scannapieco 2006) associadas às fontes de informação como elementos de metadados do tipo administrativo. Estas medidas são apuradas através de processos contínuos de avaliação da qualidade de dados e registradas no repositório de metadados.

3.3.4. Reformulador Semântico de Consultas

O Reformulador Semântico de Consultas é o componente que interage com o Gerente de Acesso a Ontologias e o Gerente de Acesso a Metadados durante a busca por fontes de informação. O processo começa com a busca por recursos nas ontologias de domínio, como detalhado no item 3.3.4.1. A seguir, o usuário escolhe dentre os recursos sugeridos aqueles que devem ser utilizados na busca. O passo seguinte, descrito no item 3.3.4.2, é a busca por fontes de informação no Repositório de Metadados através de duas consultas criadas pelo Reformulador Semântico de Consultas. O resultado de ambas as consultas é ordenado pelo componente de

acordo com as características de qualidade de dados das fontes de informação indicadas pelo usuário, conforme explicado no item 3.3.4.3.

3.3.4.1 Busca por Recursos na Ontologia

A lista de termos informada pelo usuário é submetida a processos de remoção de *stopwords*, normalização, *stemming* e/ou *lemmatization* (Manning *et al.* 2009) e a lista de termos resultante é usada para realizar uma busca textual nos rótulos dos recursos das ontologias de domínio. Nesta etapa, a interação é realizada somente com o Gerente de Acesso a Ontologias que recebe cada um dos termos para realizar a busca. O casamento entre os termos e os rótulos pode ser total ou parcial, por exemplo, ao utilizar o temo "***Aluno***" é possível recuperar os conceitos "***Aluno***", "***Aluno*** *de Graduação*", "***Aluno*** *de Mestrado*", "***Aluno*** *Avulso*" e "*Ex-****aluno***". A lista de termos também pode conter expressões entre aspas onde os termos são ligados por uma preposição, neste caso não é realizada a remoção de *stopwords*, pois a expressão pode perder o sentido ou modificar o seu significado.

A arquitetura utiliza todos os rótulos atribuídos a um recurso uma vez este pode possuir mais de um rótulo associado, esta situação ocorre quando termos sinônimos são usados para designar um mesmo recurso ou quando um recurso é conhecido por uma sigla e por seu nome por extenso. A palavra "***Aluno***", por exemplo, pode ser usada para designar o mesmo conceito que a palavra "***Estudante***". A arquitetura não faz uso de dicionário de sinônimos e com isso restringe a relação de sinonímia ao domínio e à terminologia representados pela ontologia.

O casamento entre os termos da consulta e os recursos da ontologia é o ponto de partida para o processo de expansão semântica, por isso esta etapa tem grande influência no resultado final da busca (Bhogal *et. al.* 2007). O resultado desta busca textual utilizando os rótulos é um subconjunto de recursos composto por seus rótulos, URIref e a identificação do tipo de recurso (conceito, relacionamento, atributo e instância) e, para cada recurso recuperado, é realizada uma busca com propagação por recursos vizinhos na ontologia.

O Reformulador Semântico de Consultas trata a ontologia como um grafo onde cada recurso é um nó e a ligação entre estes nós (arestas) representa a existência de uma associação entre os recursos da mesma ontologia e de ontologias previamente alinhadas. As regras de propagação da busca são definidas de acordo com o tipo de recurso para recuperar um conjunto de recursos vizinhos que estão a uma distância semântica ***d*** definida pelo número de arestas que ligam os recursos. A relação de equivalência entre recursos de ontologias diferentes representada pelos alinhamentos formais é um tipo especial de associação, a sua distância semântica é igual a 0 pois

se tratam do mesmo recurso. A tabela 3.1 exemplifica um conjunto de regras de propagação para a distância semântica máxima é igual a 2.

Tabela 3.1 Regras de Propagação na Busca por Recursos da Ontologia

Tipo do recurso na Ontologia	Distância Semântica *(d)*	Regra de Propagação
Atributo	0	Atributos equivalentes
	1	Conceitos associados
	1	Conceitos associados aos atributos equivalentes
	2	Conceitos filhos que herdam o atributo
Relacionamento de domínio	0	Relacionamentos de domínio equivalentes
	1	Conceitos associados
	1	Conceitos equivalentes aos conceitos associados
	1	Conceitos associados aos relacionamentos de domínio equivalentes
Conceitos	0	Conceitos equivalentes
	1	Atributos associados
	1	Atributos de conceitos equivalentes
	1	Relacionamentos de domínio associados
	1	Conceitos pais (1 nível acima)
	1	Conceitos filhos (1 nível abaixo)
	1	Conceitos pais (1 nível acima) de conceitos equivalentes
	1	Conceitos filhos (1 nível abaixo) de conceitos equivalentes
	1	Conceitos todo, se o conceito for parte em uma relação todo-parte
	1	Conceitos partes, se o conceito for todo em uma relação todo-parte
	2	Conceitos associados através de relacionamentos de domínio
	2	Conceitos irmãos (no mesmo nível)
	2	Atributos herdados de conceitos pais
	2	Conceitos que representam as outras partes do conceito todo, se o conceito for parte em uma relação todo-parte
Instância	1	Conceito associado

O resultado desta busca são fragmentos da ontologia, cujo ponto central é o recurso recuperado pela busca textual. A figura 3.4 ilustra um fragmento recuperado pela busca a partir do conceito "***Aluno de Pós Graduação Stricto Sensu***", este é o recurso central. A regra "Conceitos associados através de relacionamentos de domínio" recupera o conceito "***Orientador***", que está a uma distância semântica igual a 2 em relação ao recurso central. Estes fragmentos são compostos por rótulos, URIref e a identificação do tipo de associação dos recursos em relação ao recurso central, além dos rótulos, URIref e a identificação do tipo do próprio recurso central.

Para superar as dificuldades que o usuário possa ter ao navegar em ontologias com muitos recursos durante a busca, a arquitetura utiliza estes fragmentos como recursos candidatos a expansão, restringindo a navegação do usuário aos recursos da

ontologia que estão no máximo a uma distância *d* da lista de termos inicial (Bhogal *et. al.* 2007). As regras de propagação podem ser aplicadas recursivamente até recuperar todos os recursos e a distância máxima atingida depende do tamanho das ontologias de domínio. Porém, apesar da garantia da completeza computacional e decidibilidade da linguagem OWL-DL, é desejável que o usuário tenha controle da expansão a ser realizada para evitar que a consulta se distancie da intenção original de busca, problema conhecido como *query drift* (Thiagarajan *et. al.* 2008).

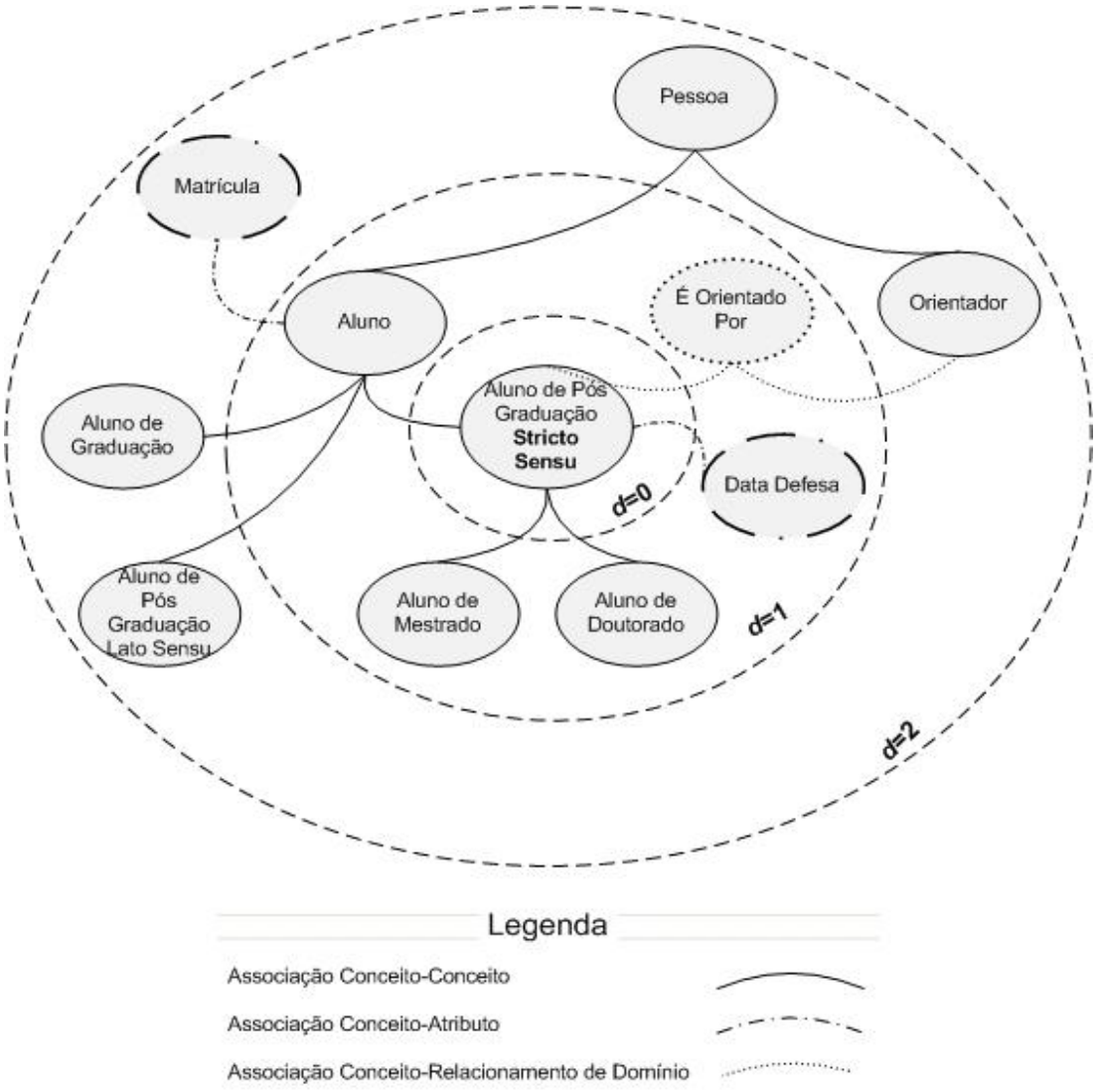


Figura 3.4 Fragmento de Ontologia Recuperado com a Busca com Propagação

Em função disto, a arquitetura utiliza uma abordagem de expansão assistida pelo usuário, onde este controla a expansão das consultas executadas no Repositório de Metadados. O resultado desta etapa de propagação é submetido à avaliação do usuário, para que o mesmo possa incluir ou não itens de acordo com os seus requisitos e realizar a desambiguação dos termos ambíguos, se houverem, a partir dos

recursos recuperados. Com isto, o usuário pode esclarecer o contexto da busca e melhorar a especificação da sua necessidade de informação, selecionando os recursos que julgar relevantes, além de adquirir conhecimento sobre o domínio e o vocabulário utilizado. Os rótulos e URIref dos recursos selecionados pelo usuário neste passo são utilizados na próxima etapa busca.

3.3.4.2 Busca por Metadados no Repositório de Metadados

A busca por metadados no repositório de metadados é realizada através de duas consultas: (1) busca utilizando os URIrefs para recuperar fontes de informação usando anotação semântica e (2) busca textual nos metadados descritivos, usando a lista semanticamente expandida de palavras-chave.

As anotações semânticas são úteis para encontrar com precisão todas as fontes de informação que contém instâncias de um determinado conceito, para determinar todos os conceitos presentes em uma fonte de informação específica e para facilitar a integração de fontes de informação através da identificação da similaridade e do relacionamento semântico entre os conceitos aos quais estas fontes de informação estão vinculadas.

As anotações também são utilizadas para identificar a subsunção entre diferentes fontes de informação. A subsunção (do inglês *subsumption*) caracteriza o relacionamento de especialização e generalização entre recursos. Desta forma, se o usuário selecionar o conceito **#Professor** para realizar a busca no Repositório de Metadados, as fontes de informação associadas ao conceito **#Professor_Adjunto** também são recuperadas pois **#Professor_Adjunto** é um conceito filho de **#Professor**.

Quando a busca é realizada somente nas anotações semânticas, o resultado melhora a precisão, pois retorna menos falsos positivos. Por outro lado, algumas fontes de informações relevantes não serão recuperadas por não terem sido anotadas com todos os recursos devidos, o que reduz a cobertura do resultado.

Para compensar esta limitação, também é realizada uma busca textual nos metadados descritivos selecionados em cada esquema de metadados. A lista de palavras utilizada nesta busca é a lista original expandida com novos termos que correspondem aos rótulos dos recursos selecionados pelo usuário na etapa anterior. Antes de serem acrescentados, os rótulos também são submetidos a processos de remoção de *stopwords*, normalização, *stemming* e/ou *lemmatization*.

Os termos adicionados são semanticamente relacionados com os termos originais, pois pertencem aos domínios que foram modelados através das ontologias utilizadas para busca e anotação. Desta forma é possível reduzir o número de

metadados de fontes de informação relevantes que não foram recuperados com a busca semântica nas anotações e aumentar a cobertura do resultado. A busca textual utilizando palavras-chave compensa a falta de completeza do processo de anotação. Desta forma, uma busca completa o resultado da outra.

O passo de busca com propagação realizado na primeira etapa da busca e as duas abordagens de expansão de consultas realizadas nesta etapa permitem a recuperação de metadados de fontes de informação que são relevantes para atender a necessidade de informação, apesar de não conter exatamente os termos especificados na lista inicial (Crestani 1997) no conteúdo de seus elementos descritivos ou não terem sido anotadas com os recursos mapeados diretamente por estes termos.

O resultado de buscas por fontes de informação em um repositório de metadados usando consultas por palavras-chave e uma lista de URIrefs não é um conjunto exato de respostas, por isso é necessário aplicar uma regra de ordenação. Na arquitetura proposta, a ordem dos resultados é dependente do contexto, pois está relacionada com as características de qualidade que estas fontes possuem e os requisitos de qualidade que atendem à necessidade de informação.

3.3.4.3 Tratamento do Resultado da Busca por Metadados

A qualidade dos dados é uma preocupação em processos de integração de informações, pois dados com baixa qualidade representam um obstáculo para esforços de integração de informações (Ziegler e Dittrich 2007) e por isso este é um critério importante na seleção das fontes de informação a serem utilizadas.

De modo geral, qualidade de dados pode ser definida como “adequação ao uso”, por isso a qualidade não pode ser medida ou avaliada sem considerar o seu contexto (Tayi e Ballou 1998). Contexto em uma definição mais geral é “as condições inter-relacionadas em que alguma coisa existe ou acontece”, ou seja, depende das circunstâncias em que algo ocorre (Bhogal *et. al.* 2007).

Os critérios de qualidade para seleção das fontes são dependentes do contexto da necessidade de informação, por isso o usuário deve indicar os critérios de qualidade adequados a sua necessidade de informação e a prioridade entre eles. Deste modo, é possível informar, por exemplo, que para atender a uma necessidade de informação a consistência é mais importante que o tempo de resposta de uma fonte de informação.

O esquema de metadados deve conter elementos que permitam a especificação das características de qualidade de cada fonte. Os critérios de qualidade informados pelo usuário só poderão ser considerados para ordenação dos resultados

se estes forem coletados e registrados no Repositório de Metadados, associados às fontes de informação, através de processos de avaliação de qualidade.

3.3.5 Registro de Consultas

O sucesso do uso de ontologias para expansão de consultas depende de vários fatores e um deles é a qualidade e adequação da própria ontologia para esta função (Bhogal *et. al.* 2007). A ontologia utilizada deve ser o mais completa possível, pois caso contrário as consultas não serão aprimoradas de forma adequada, por deficiência do próprio modelo do conhecimento.

Considerando esta fragilidade, a arquitetura propõe que as consultas sejam registradas em um *log* para armazenar os termos utilizados na lista inicial de palavras-chave, os recursos selecionados para expansão de consulta e o julgamento de relevância de cada fonte realizado pelos usuários. Este *log* é usado para investigar as divergências terminológicas e conceituais entre quem possui a necessidade de informação e realiza a busca no repositório de metadados e quem desenvolve e cataloga as fontes de informações. Desta forma, a implantação da arquitetura não requer que as ontologias de domínio utilizadas sejam completas, estas ontologias podem ser aprimoradas com os termos extraídos do *log* de consultas da própria arquitetura.

Os termos informados nas consultas podem não ser mapeados em recursos da ontologia ou, mesmo quando são mapeados, estes recursos podem não ser selecionados pelos usuários. Como em um ambiente real seria necessário muito esforço humano para analisar o resultado da busca para cada termo encontrado no *log*, a análise somente dos termos mais utilizados pode ser adotada. A análise destes termos deve ser realizada por um especialista de domínio, como parte de uma metodologia de engenharia de ontologias, e pode ser apoiada por uma ontologia lingüística independente de domínio para identificar quais podem ser usados para adicionar novos rótulos para recursos que já existem ou novos recursos na ontologia. O método 101 (Noy e McGuinness 2001), por exemplo, estabelece as seguintes etapas para o processo de engenharia de ontologias: (1) determinar o domínio e escopo da ontologia; (2) considerar o reuso de ontologias existentes; (3) enumerar termos importantes na ontologia; (4) definir as classes e a hierarquia de classes; (5) definir as propriedades das classes através de relacionamentos; (6) definir as propriedades das classes através de atributos e (7) definir instâncias. Seguindo este método, os termos extraídos do *log* podem ser usados como insumos para a terceira etapa deste processo.

Além dos termos, os recursos selecionados pelos usuários durante a busca também são registrados. Se os rótulos destes recursos recuperam fontes de informação que são consideradas relevantes através da busca textual, mas que não estão anotadas com os URIrefs, os responsáveis pelas fontes devem analisar estes recursos para verificar se a anotação semântica deve ser realizada. Este processo também se aplica às novas fontes de informação, pois apesar de serem submetidas ao processo de anotação semântica no momento de sua catalogação no Repositório de Metadados, ainda podem ser anotadas com recursos que foram acrescentados na ontologia de domínio após a sua catalogação.

O Gestor do Repositório de Metadados também pode utilizar o *log* para identificar os recursos existentes nas ontologias de domínio e que não recuperam nenhuma fonte de informação (através de seus rótulos ou anotações semânticas), pois esta situação pode indicar fontes que existem na organização e não estão registradas no repositório ou fontes que ainda não foram desenvolvidas ou adquiridas.

3.4 Classificação da Arquitetura Proposta

A arquitetura apresentada pode ser classificada, a partir das características apresentadas na figura 3.5, de acordo com o esquema de categorização proposto por Mangold (2007), conforme exposto a seguir:

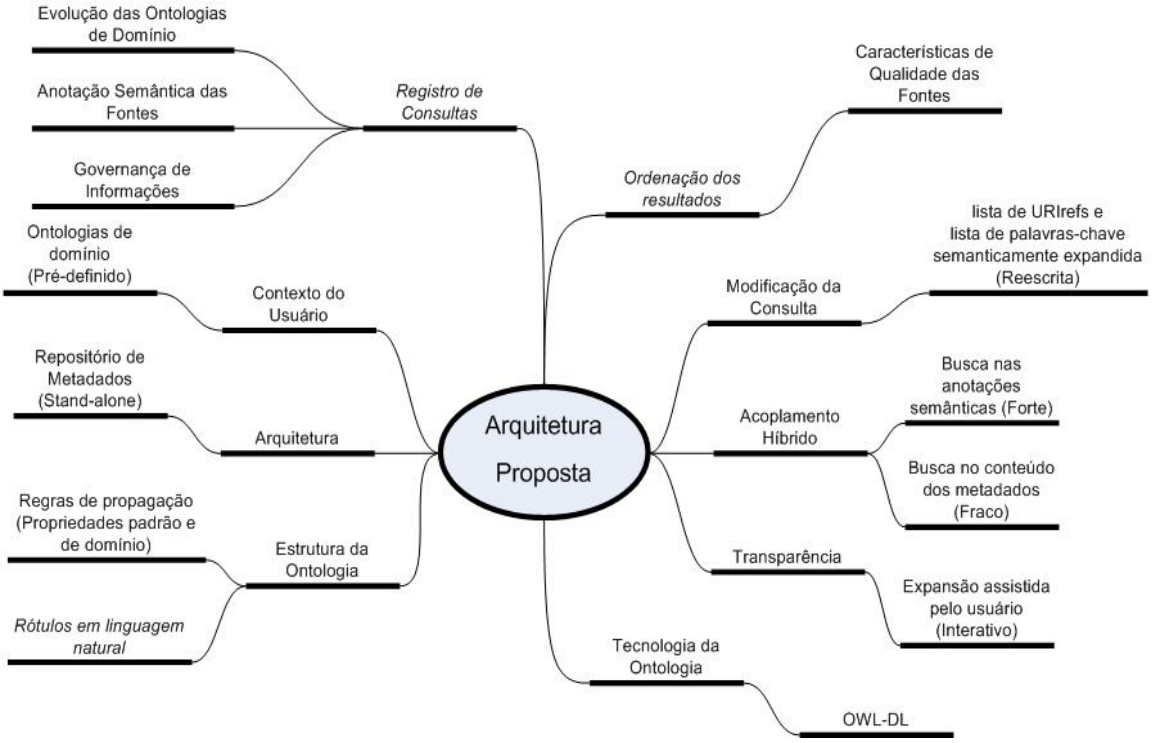


Figura 3.5 Principais Características da Arquitetura Proposta

Arquitetura *stand-alone*: a busca é restrita ao Repositório de Metadados, pois este é considerado o ponto central para registro de todas as fontes de informação relevantes através da descrição de seu conteúdo e das características de qualidade.

Acoplamento híbrido: a busca permite a recuperação de metadados de fontes de informação que foram ou não anotadas semanticamente no momento de sua catalogação. Devido a esta característica, a arquitetura pode ser aplicada para estender um sistema de registro de metadados existente.

Propriedades padrão e de domínio: a arquitetura faz uso de propriedades padrão como hierarquias e relações todo/parte assim como propriedades dependentes de domínio para recuperar fragmentos de ontologias de domínio armazenadas no Repositório de Ontologias durante a primeira etapa da busca por fontes de informação.

Contexto do usuário: as ontologias, utilizadas para anotação das fontes de informação no repositório de metadados, restringem a busca em relação aos seus domínios, tornando o contexto do usuário como pré-definido.

Transparência: através de um processo de busca em duas etapas, a arquitetura utiliza uma abordagem interativa para esclarecer a intenção de busca. A primeira etapa, além de esclarecer o significado dos termos dentro do domínio quando houver ambigüidade, permite que o usuário melhore a especificação de sua consulta e também adquira conhecimento sobre o domínio ao recuperar fragmentos da ontologia.

Modificação da consulta: o Reformulador Semântico de Consultas transforma uma lista inicial de palavras-chave em duas consultas: a primeira utiliza os URIrefs dos recursos selecionados pelo usuário e a segunda, uma lista semanticamente expandida de palavras-chave composta pela lista inicial de termos acrescentada dos rótulos destes mesmos recursos.

Tecnologia da ontologia: as ontologias de domínio devem ser representadas através da linguagem *OWL* (*Web Ontology Language*), por se tratar de uma recomendação da W3C, pela possibilidade de reuso de ontologias de domínio já disponíveis nesta linguagem, pela capacidade de associar rótulos e comentários aos recursos para apoio à interpretação humana, entre outras características. A variação OWL-DL é a mais adequada devido a sua computabilidade e expressividade.

Além disso, duas características diferenciam a arquitetura proposta das ferramentas de busca semântica analisadas por Mangold (2007): **Ordenação dos Resultados** e **Registro de Consultas**. O resultado da busca por fontes de informação é ordenado segundo as características de qualidade das fontes de informação registradas no repositório de metadados e selecionadas pelo usuário para atender a sua necessidade de informação. O *log* das consultas fornece insumos para os

processos de Anotação Semântica, Evolução das Ontologias de Domínio e Governança de Informações, integrando as atividades de busca e catalogação de fontes de informação com outras atividades dentro de iniciativas de estabelecimento da Arquitetura de Informações.

3.5 Considerações Finais sobre a Proposta

A arquitetura apresentada neste capítulo têm com objetivo tornar mais eficiente o resultado da busca por fontes de informação, pois explora o conhecimento do domínio e permite que o usuário especifique os conceitos envolvidos em sua necessidade de informação de uma forma mais explícita, precisa e baseada em sua intenção de busca. Nesta arquitetura, a expansão da consulta assistida pelo usuário minimiza os problemas de divergência terminológica e o uso de ontologias acrescenta semântica formal nos metadados descritivos. Os conflitos de nomenclatura e estrutura das fontes de informação são tratados durante a busca através da utilização de mais de um rótulo por recurso e das anotações semânticas, independente de estrutura interna das fontes de informação.

Para garantir que todas as fontes de informação relevantes que suportam os processos de negócio, sejam elas internas ou externas a organização, estejam registradas no repositório de metadados, é necessário que políticas e procedimentos organizacionais para registro destas sejam definidos, seguidos e monitorados. A governança da informação promove o estabelecimento de processos de gestão do acervo de informações ao longo de todo o seu ciclo de vida e o repositório de metadados corporativos é um dos recursos que apóiam estes processos.

No próximo capítulo será apresentado um protótipo da arquitetura proposta construído para ser utilizado nas fases de avaliação assim como os softwares utilizados.

4. Protótipo

Para avaliar a arquitetura proposta, através de um experimento e um estudo de caso, foi construído um protótipo chamado SEM-SII [1]. Este capítulo descreve os softwares utilizados e suas funcionalidades assim como o detalhamento do protótipo.

Para realizar a busca textual por recursos da ontologia e por fontes de informação usando os metadados descritivos que foram armazenados no SGBD *PostgreSQL* foi utilizada a funcionalidade conhecida como *Full Text Search* que será apresentada na próxima seção. A navegação nas ontologias foi suportada pelo *software ontology browser* apresentado na seção 4.2. O esquema de metadados da proposta de Py *et. al.* (2009) é apresentado na seção 4.3 e a seção 4.4 apresenta a interface gráfica e as funcionalidades de catalogação e busca por fontes de informação utilizando este esquema. A funcionalidade de busca apresentada foi utilizada pelos participantes durante o experimento. A aplicação da prova de conceito do estudo de caso também utilizou a mesma funcionalidade de busca instanciada para um novo esquema de metadados, o do portal *DadosGov* COI-PR.

4.1 *PostgreSQL* e *Full Text Search*

O SGBD *PostgreSQL* 8.4 [2] foi utilizado no experimento por se tratar de um software livre, de fácil instalação e utilização e também por contar com uma funcionalidade, chamada *Full Text Search* [3], que permite a recuperação de conteúdo textual em linguagem natural que satisfaz uma consulta especificada também em linguagem natural. Durante o estudo de caso, seu uso foi uma escolha natural, pois os metadados e dados do portal *DadosGov* COI-PR já estavam armazenados neste SGBD.

Full Text Search permite que qualquer campo do tipo texto que pertence a uma tabela ou a concatenação de campos de uma ou mais tabelas sejam tratados como

[1] http://semsii.uniriotec.br/semsii/

[2] http://www.postgresql.org/docs/8.4/static/index.html

[3] http://www.postgresql.org/docs/8.4/static/textsearch.html

um documento para as operações de busca textual e indexação de texto. Esta funcionalidade suporta as operações de normalização, remoção de *stopwords*, *stemming* e *lemmatization*. Estas operações foram aplicadas na indexação dos rótulos dos recursos da ontologia e dos elementos descritivos do repositório de metadados e na lista de termos das consultas realizadas para busca por fontes de informação.

Cada ontologia foi convertida no formato *RDF / OWL Database* através do editor de ontologias *Protégé* 3.4.3 e armazenada no banco. Os elementos descritos selecionados dos esquemas de metadados foram submetidos a uma etapa de pré-processamento, utilizando a função *to_tsvector* para converter seu conteúdo para o tipo de dados *tsvector,* cujo armazenamento é otimizado para busca textual. A primeira etapa deste processo de conversão é a separação do conteúdo em *tokens* que são submetidos aos dicionários de acordo com o seu tipo.

O mapeamento das diferentes variações das palavras em relação ao seu radical é feito com apoio de dicionários como o *Ispell* [4] e o *Hunspell* [5] ou através de regras de remoção de sufixos e prefixos como o *Snowball* [6]. De modo geral, um dicionário é um programa que recebe um *token* e retorna um conjunto de um a vários lexemas, um conjunto vazio se este *token* for uma *stopword* ou NULO se o *token* não for reconhecido. Estes dicionários possuem arquivos de configuração que devem ser armazenados no servidor usando a codificação de caracteres UTF-8, independente do padrão de codificação configurado para o banco de dados. A configuração do conjunto de dicionários deve partir do mais específico até o mais geral. Alguns dicionários padrão são fornecidos junto com a instalação do *software* e estes podem ser atualizados e novos dicionários podem ser adicionados para atender a necessidades particulares. Um dos arquivos de configuração que apóia estas operações contém a lista de *stopwords*. Esta lista de *stopwords* pode ser atualizada extraindo as palavras com maior freqüência em uma coleção de documentos. A função *ts_stat* dá suporte a este processo e retorna cada palavra existente na coleção, o número de documentos e o número total de vezes que esta ocorre.

As configurações utilizadas durante o experimento e o estudo de caso estão especificadas no anexo I (Configuração FTS). O arquivo de *stopwords* em português criado com a instalação foi alterado para incluir palavras, identificadas através da função *ts_stat*, encontradas com muita freqüência no conteúdo dos metadados. Os

[4] http://ficus-www.cs.ucla.edu/geoff/ispell.html

[5] http://hunspell.sourceforge.net/

[6] http://snowball.tartarus.org/

demais arquivos de configuração correspondem a dicionários em português para o *Snowball (portuguese_stem* [7]*)*, *Ispell* (*pt_ispell* [8]*)* e *Hunspell* (*pt_hunspell* [9]*).*

A realização do casamento dos termos da consulta com os rótulos dos recursos e com o conteúdo dos metadados depende da conversão da lista de termos no tipo de dados *tsquery* realizada através de funções (*to_tsquery* e *plainto_tsquery*). Os operadores permitidos na especificação da consulta são: & (AND), | (OR) e ! (NOT) e estes podem ser combinados separando as expressões entre parentesis. O operador de casamento entre a consulta (*tsquery*) e o conteúdo textual (*tsvector*) é o @@. O comando apresentado abaixo exemplifica a recuperação dos URIrefs e rótulos dos recursos que possuem um rótulo que permite o casamento, parcial ou totalmente, com termo professor.

```
SELECT distinct onto.protege.frame AS uriref, short_value AS rotulo
FROM onto.protege
WHERE protege.frame_type = 9 and protege.slot =
'http://www.w3.org/2000/01/rdf-schema#label'  and
to_tsvector(short_value) @@ to_tsquery('professor');
```

O resultado deste casamento pode ser ordenado de acordo com a freqüência, a proximidade dos termos no seu conteúdo e a importância da seção onde estes ocorrem, caso tenham sido atribuídos pesos às seções. Uma função (*ts_rank*) calcula a pontuação do casamento e este cálculo pode ser diferenciado de acordo com o número de palavras que contém, conforme mostra a tabela 4.1.

Tabela 4.1 Opções de Cálculo da Pontuação pela Função *ts_rank*

Parâmetro	Regra de cálculo
0 (default)	Ignorar o tamanho do documento
1	Divide a pontuação por 1 + log do tamanho do documento
2	Divide a pontuação pelo tamanho do documento
8	Divide a pontuação pelo número de palavras distintas no documento
16	Divide a pontuação por 1 + log do número de palavras distintas no documento
32	Divide a pontuação por ela mesma + 1

Antes de exibir o conteúdo completo recuperado, é possível apresentar algumas partes deste para destacar a ocorrência dos termos da consulta dentro de trechos do documento através da função *ts_headline*. No exemplo abaixo, são formatados fragmentos de no mínimo 15 e no máximo 35 palavras onde os termos são

[7] http://snowball.tartarus.org/algorithms/portuguese/stemmer.html

[8] http://www.ime.usp.br/~ueda/br.ispell/

[9] http://ftp.services.openoffice.org/pub/OpenOffice.org/contrib/dictionaries/pt_PT-pack.zip

destacados entre os delimitadores ** e *,* produzindo um efeito de negrito e itálico se visualizados em um browser pois se tratam de *tags* em HTML.

```
ts_headline(docto, q, 'MaxWords=35, MinWords=15, StartSel=<b><i>, StopSel=</b></i>')
```

Esta funcionalidade dispõe de dois tipos de índices para otimizar as consultas com o operador @@: GiST (*Generalized Search Tree*) e GIN (*Generalized Inverted Index*). No GiST, a coluna indexada pode ser *tsvector* ou *tsquery*; a atualização deste índice é mais rápida que o GIN e apresenta melhor desempenho nas consultas caso o número de lexemas distintos seja menor que 100.000. Já no GIN, a coluna deve ser do tipo *tsvector* e o tempo de resposta da consulta pode ser até três vezes mais rápido que no GiST, variando de acordo com o número de palavras únicas, mas o tempo para criação do GIN é três vezes maior e o tamanho de duas a três vezes maior. Como regra geral, segundo a documentação do software, o GIN é mais indicado para coleção de documentos estáticos e GiST para dinâmicos.

4.2 *Ontology-Browser*

Para visualizar as ontologias de domínio utilizadas no experimento e no estudo de caso foi usado o software *ontology-browser.* Este *software* foi escolhido por oferecer a busca por recursos de uma ontologia em OWL através de seus rótulos, a expansão dos nós na forma de uma lista indentada (à esquerda), o detalhamento dos conceitos, relacionamentos (padrão e de domínio), atributos e instâncias a partir de um recurso (à direita) e a navegação entre os recursos utilizando hiperlinks em uma interface web conforme apresentado na figura 4.1. O software foi desenvolvido na Universidade de Manchester, a versão 1.4.2 está disponível para download a partir do repositório *Google Code* [10] e utiliza a OWL API [11]. Esta API, escrita em Java, permite a criação, manipulação e serialização de ontologias em OWL e é usada na versão 4.0 do Protégé.

O projeto disponibiliza também um servidor na web onde é possível carregar uma ou várias ontologias e explorá-las utilizando o *ontology-browser* [12]. O *ontology-browser* também pode ser disponibilizado em um servidor próprio, e para isto é necessária a instalação de um servidor Java para aplicações web como, por exemplo,

[10] http://code.google.com/p/ontology-browser/

[11] http://owlapi.sourceforge.net/

[12] http://owl.cs.manchester.ac.uk/browser/

o *Tomcat*, e a publicação da aplicação a partir do arquivo *war* disponibilizado para *download*.

A carga de uma ontologia pode ser feita interativamente especificando o seu endereço na web através da opção *Load Ontologies* ou através da opção */ontologies* na URL do servidor com os parâmetros *action*, *clear* e *uri*. O parâmetro *uri* especifica a localização física do arquivo .owl, *action* determina a ação a ser realizada, neste caso *load*, e o parâmetro *clear=true* sobrepõe todas as ontologias que poderiam ter sido carregadas anteriormente na mesma sessão. Outras opções para o parâmetro *action* são *remove*, para remover a ontologia carregada, e *reload*, para carregar uma nova versão. Depois de carregada, é possível obter um link permanente para referenciar a sessão na qual a ontologia foi carregada. O servidor do projeto foi utilizado no experimento mas durante o estudo de caso, em função de problemas de desempenho neste servidor, foi necessário utilizar um servidor próprio.

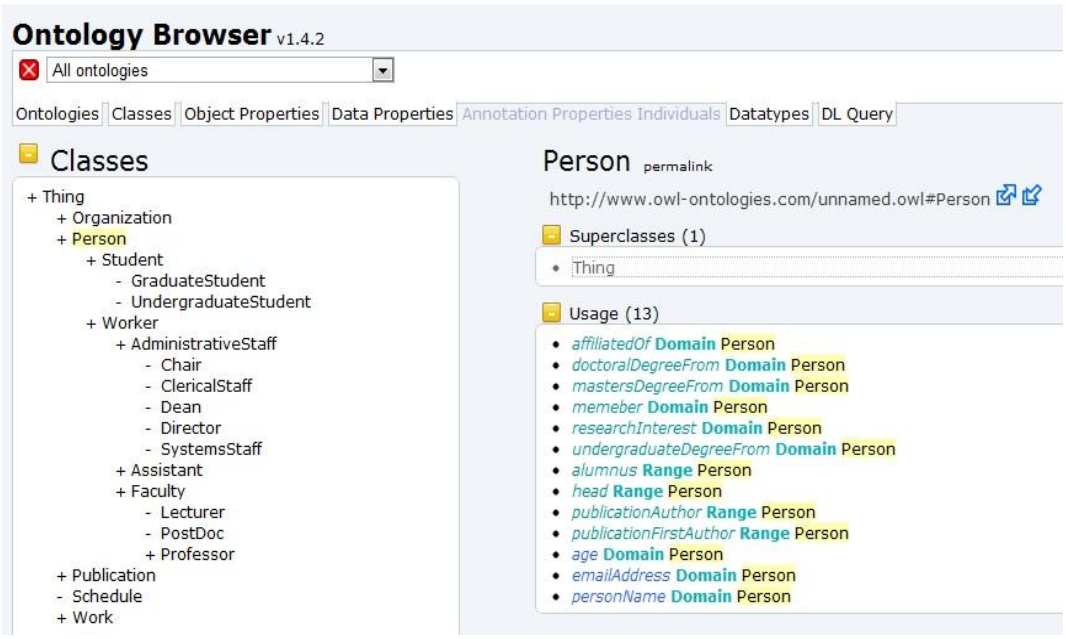


Figura 4.1 – Visualização da Hierarquia de Classes e dos Recursos Associados a uma Classe

A busca por recursos em uma ontologia pode ser realizada interativamente a partir de parte de seu identificador informado na caixa de texto ao lado do botão *Find* ou através da opção */find* na URL do servidor com os parâmetros *type* e *input*. Os tipos podem ser recursos (*entities*), conceitos (*classes*), relacionamentos de domínio (*objectproperties*), atributos (*dataproperties*), instâncias (*individuals*) e ontologias (*ontologies*). A interface também permite filtrar os recursos de acordo com o tipo. Usando o identificador completo do recurso no parâmetro uri e opção entities/ na URL são exibidos os demais recursos com que este se relaciona.

4.3 Esquema de Metadados

Py *et. al.* (2009) propõem um método com passos detalhados para identificação de conceitos e fontes de dados durante um processo de integração de dados. O método se baseia na análise de glossários em modelos de processos de negócio e assume que todos os processos possuem uma única identificação e descrição para uma entidade. O método é suportado por um framework que utiliza serviços além de um repositório de metadados armazenado em um banco relacional cujo modelo lógico é apresentado na figura 4.2.

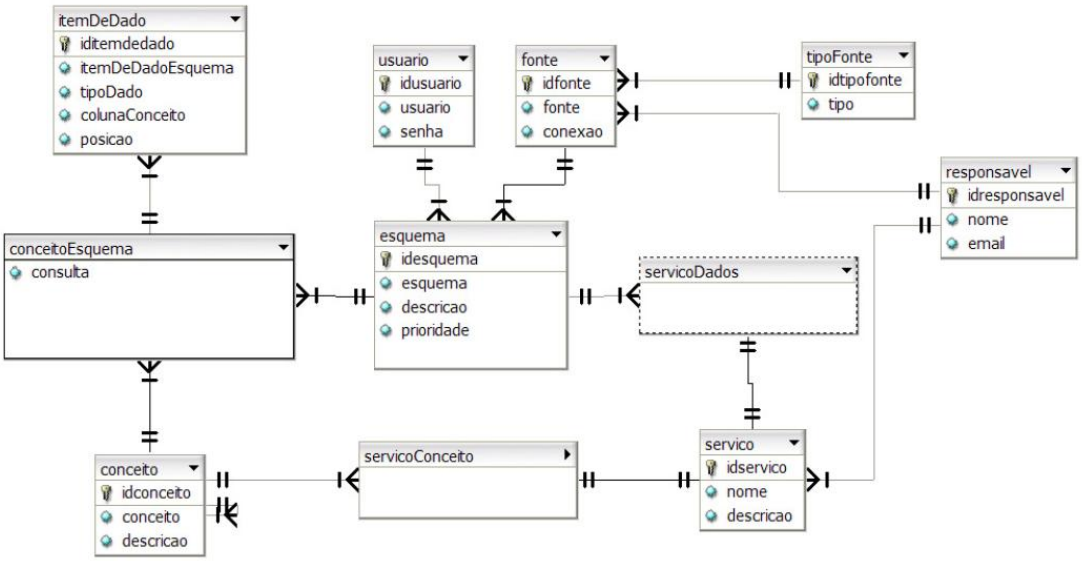


Figura 4.2 – Modelo Lógico do Repositório de Metadados de Py *et. al.* (2009)

O esquema de metadados foi analisado para identificar os elementos descritivos e as características de qualidade das fontes de informações. O resultado desta análise é apresentado no capítulo 5 (item 5.1.2). Os elementos descritivos selecionados do esquema de metadados foram submetidos a normalização, remoção de *stopwords*, *stemming* e *lemmatization* e o resultado foi armazenado em uma coluna do tipo *tsvector* indexada utilizando um índice GiST.

4.4 Detalhamento do Protótipo

A interface gráfica agregada ao *framework* possui três opções: **Buscar**, **Visualizar Ontologia** e **Registrar Metadados**.

A opção '**Visualizar Ontologia'** permite explorar as ontologias de domínio selecionadas do repositório de ontologias utilizando o *ontology-browser*.

A catalogação das fontes de informação é realizada através da opção '**Registrar Metadados'**. A figura 4.3 apresenta uma das telas de catalogação onde é

realizado o cadastramento dos conceitos de dados e serviços de conceitos e a anotação semântica da fonte de informação.

Figura 4.3 – Catalogação de Conceito de Dados e o Serviço de Conceito

As consultas são realizadas através da opção '**Buscar**' e visam recuperar as fontes de informação que atendem a uma necessidade de informação em particular. A interface de busca por palavras-chave, apresentada na figura 4.4, aceita até 6 termos por consulta e a especificação de um operador para combinação dos termos como TODOS os termos (E) ou QUALQUER termo (OU). Esta lista de palavras-chave é submetida à normalização, remoção de *stopwords*, *stemming* e *lemmatization* e cada termo resultante é usado para realizar uma busca textual por recursos na ontologia através de seus rótulos utilizando as funcionalidades do *Full Text Search*.

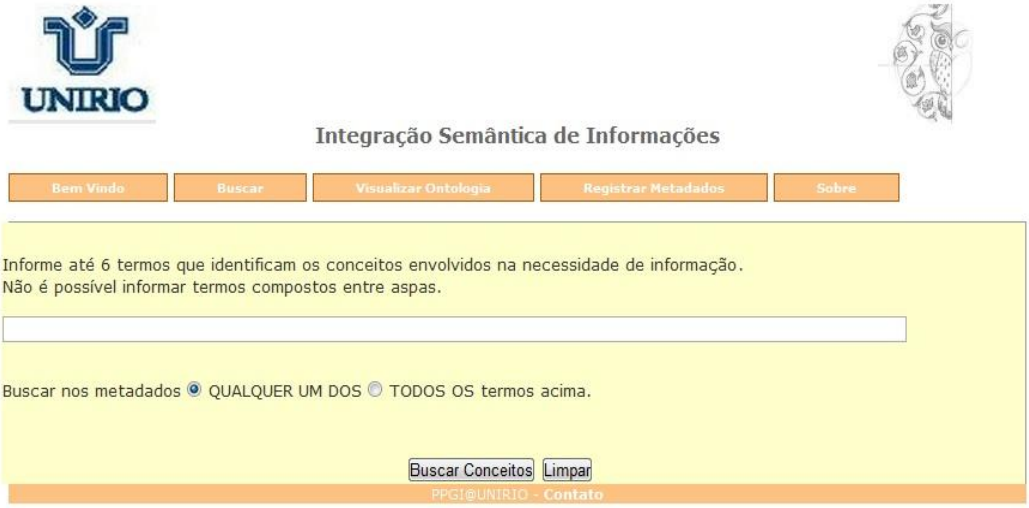


Figura 4.4 – Interface de Busca por Palavras-chave

Para cada recurso identificado na ontologia, podendo ser por casamento total ou parcial de termos, são recuperados o URIref e a lista de rótulos e é iniciada a busca com propagação por recursos associados. Em caso de classes (*owl:Class*), esta etapa retorna as classes equivalentes (*owl:equivalentClass*), classes filhas, classes pais (*rdfs:subClassOf*), classes irmãs, classes todo, classes parte (*owl:unionOf*), relacionamentos (*owl:ObjectProperty)* e atributos (*owl:DatatypeProperty).* Se o casamento recuperar um relacionamento ou atributo (*rdf:Property*) então são retornados os relacionamentos e atributos equivalentes (*owl:equivalentProperty*) e as classes relacionadas (*rdfs:domain* / *rdfs:range*). A distância semântica máxima atingida com estas regras de propagação é 2. O resultado final desta etapa de busca são fragmentos da ontologia que são apresentados ao usuário agrupados por palavra chave, como é mostrado na figura 4.5. O usuário analisa os recursos e adiciona na consulta aqueles que melhoram a especificação de sua consulta.

Os URIrefs dos recursos selecionados são utilizados para realizar a busca nas anotações. As fontes anotadas com pelo menos um dos recursos selecionados, ou com um conceito filho de algum destes (*is-a*) ou com instâncias de um conceito (*instace-of*) são recuperadas. Já os rótulos são adicionados à lista inicial de palavras-chave para montar a lista de palavras-chave semanticamente expandida e uma busca textual nos elementos descritos selecionados é executada utilizando os recursos do *Full Text Search*. Por exemplo, a partir dos termos "**Aluno**" **E** "**Mestrado**" e da seleção do conceito **#Mestrando**, que possui os rótulos "***Aluno de Mestrado***" e "***Mestrando***", é criada a lista de palavras-chave semanticamente expandida *"aluno mestrado" OU "mestrando"* e gerada a consulta *((aluno & mestrado)* | *mestrando).*

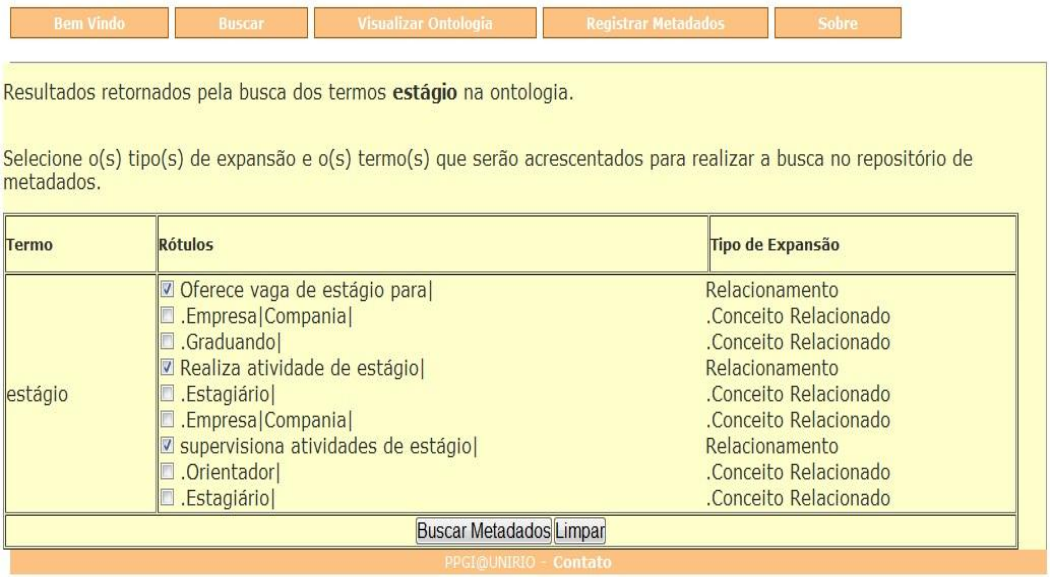


Figura 4.5 – Recursos Recomendados para Expansão da Consulta

Para cada fonte de informação recuperada são apresentados seus metadados e a opção de visualização do seu modelo de dados e do conceito da ontologia, quando tiver sido anotado. O usuário deve analisar estas informações para avaliar se a fonte de informação é relevante. Na figura 4.6 são apresentadas duas fontes de informação candidatas recuperadas a partir da consulta pelo termo "***Estágio***" no repositório de metadados utilizado no experimento. A primeira possui anotação semântica, pois a opção "**Visualizar anotação semântica**" está habilitada.

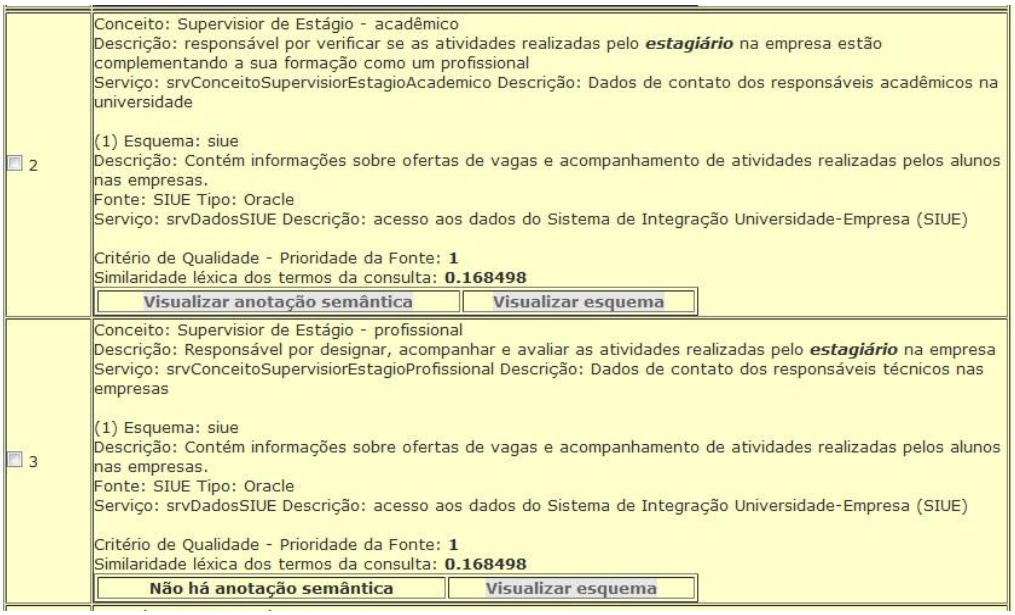


Figura 4.6 – Metadados das Fontes de Informação Recuperadas

4.5 Considerações Finais sobre o Protótipo

O protótipo apresentado neste capítulo agregou ao *framework* Py *et. al.* (2009) uma interface gráfica para catalogação e busca por fontes de informação no repositório de metadados, além de alterações no esquema de metadados para permitir a anotação semântica das fontes de informação catalogadas.

No próximo capítulo será apresentado o experimento e no capítulo 6 o resultado do estudo de caso. Estes métodos foram utilizados para avaliar a arquitetura proposta com apoio do protótipo construído.

5. Avaliação da Proposta - Experimento

Um experimento foi utilizado na primeira parte da avaliação da arquitetura proposta com o objetivo de comparar a a eficiência do resultado das consultas realizadas no repositório de metadados com o apoio da arquitetura proposta em relação a outras abordagens. O resultado esperado é que a busca semântica obtenha melhor desempenho, uma vez que esta explora o conhecimento do domínio e permite que o usuário especifique de maneira mais precisa a sua intenção durante a atividade de busca por fontes de informação. Este método foi escolhido por permitir a análise quantitativa das variáveis dependentes (precisão, cobertura e medida F) que determinam o desempenho da abordagem a partir da manipulação de variáveis independentes (número de consultas executadas, termos informados pelo usuário, recursos selecionados pelo usuário). Ao final do experimento, também foi aplicado um questionário aos participantes para coletar informações qualitativas quanto à utilização da ferramenta, o perfil dos mesmos e do ambiente organizacional onde estes estão inseridos.

Este capítulo começa apresentando na seção 5.1 a dinâmica do experimento. Os resultados do experimento e as abordagens de comparação são apresentados na seção 5.2. Em seguida, a seção 5.3 apresenta as respostas do questionário enviado aos participantes e a seção 5.4 finaliza o capítulo com as considerações finais sobre esta fase da avaliação.

5.1 Experimento

Os experimentos encontrados na literatura para recuperação da informação utilizam com freqüência as medidas de cobertura e precisão para avaliar o desempenho das abordagens do ponto de vista da qualidade de seu resultado. O cálculo destas medidas requer que um especialista identifique dentro da coleção de documentos quais são relevantes, ou seja, atendem uma determinada necessidade de informação.

Uma coleção de referência muito utilizada em experimentos para avaliação e comparação de abordagens de recuperação de informação é o TREC (*Text Retrieval Conference*). Além dos conjuntos de documentos de domínios específicos, esta coleção também contém as descrições das necessidades de informação associadas a um tópico e o julgamento de relevância em cada uma delas (Manning *et. al.* 2009). As consultas podem ser derivadas automática ou manualmente a partir do conjunto de tópicos selecionado.

Não foi encontrada na literatura uma coleção de referência, como o TREC, aplicada a abordagens de busca semântica. Tal coleção, além dos documentos, necessidades de informação e julgamento de relevância também deveria conter as ontologias de referência e a ligação entre os recursos destes e os documentos da coleção através de anotação semântica.

O projeto SEALS (*Semantic Evaluation At Large Scale*) está em processo de definição de uma metodologia e disponibilização da infraestrutura necessária para avaliação de abordagens de busca semântica (Wrigley *et. al.* 2010). A avaliação considera aspectos de desempenho como consumo de CPU e memória e por isso a plataforma SEALS fornece o ambiente de hardware necessário para execução dos experimentos através da virtualização de servidores na maioria dos sistemas operacionais.

Porém, dois aspectos particulares tornam este tipo de avaliação mais complexa: a metáfora de busca, como palavras-chave, linguagens específicas e navegação em grafos, e o comportamento do usuário. Com isto, duas fases de avaliação estão previstas, a primeira fase é automática para determinar medidas como precisão e cobertura e a segunda é interativa para mensurar aspectos de usabilidade, sendo aplicado inclusive um questionário ao final desta fase. Na primeira campanha de avaliação, SEALS 2010, somente as ferramentas de recuperação de ontologias foram avaliadas. Conjuntos de dados mais complexos como documentos não estruturados ou semi-estruturados serão considerados nas próximas avaliações do projeto.

Coleções de referência não foram utilizadas para avaliação da arquitetura desta proposta em função de duas premissas: (1) um esquema de metadados com elementos descritivos e de qualidade das fontes de informação e (2) uma ontologia de domínio que representasse os conceitos contidos nestas fontes. Assim, a preparação do ambiente para realização do experimento requereu:

1) Uma ontologia de domínio que modelasse o contexto administrativo de uma instituição de educação superior. A ontologia utilizada é apresentada no item 5.1.1.

2) A escolha de um esquema de metadados e análise deste esquema para identificar os elementos descritivos e os elementos referentes à qualidade. O esquema de metadados escolhido e o resultado desta análise são descritos na seção 5.1.2.

3) A construção do protótipo, chamado SEM-SII e apresentado no capítulo 3, com as funcionalidades especificadas pela arquitetura e sua disponibilização na internet.

4) A descrição e anotação semântica das fontes de informação dos sistemas de apoio administrativo de uma instituição de educação superior. Os sistemas são apresentados na seção 5.1.3.

5) A elaboração de três necessidades de informação e identificação das fontes de informação relevantes associadas a cada uma. As necessidades de informação e as fontes selecionadas, que estão elencadas na seção 5.1.4.

5.1.1 Ontologia de Domínio

A ontologia de domínio utilizada na busca e anotação dos metadados deve descrever os conceitos, relacionamentos e atributos específicos do contexto. Como não foi encontrada uma ontologia de referência que já atendesse a este critério, duas opções foram consideradas: a modelagem de uma nova ontologia ou o reuso de uma ontologia existente.

Uma ontologia sobre o contexto administrativo de uma instituição de ensino superior foi encontrada em (Yuan *et. al.* 2006) e serviu para avaliar uma ferramenta, chamada MapOnto, que permite a descoberta semi-automática de mapeamentos entre diferentes modelos de dados e ontologias. Esta ontologia, parcialmente apresentada na figura 5.1, contém 55 conceitos, 27 relacionamentos, 12 atributos e nenhum rótulo e foi utilizada como ponto de partida para modelagem da ontologia utilizada no experimento.

Cada recurso desta ontologia foi analisado para identificar se era aplicável ao contexto administrativo de uma instituição de educação superior no Brasil. Novos conceitos, relacionamentos e atributos foram incluídos de modo a refletir os dados existentes nos sistemas de informação que suportam os processos administrativos dentro deste contexto. Para fins de ilustração das alterações realizadas, serão citados alguns exemplos e, por questões de simplificação do texto, foi convencionada a utilização somente do fragmento do URIref para referenciar os recursos da ontologia.

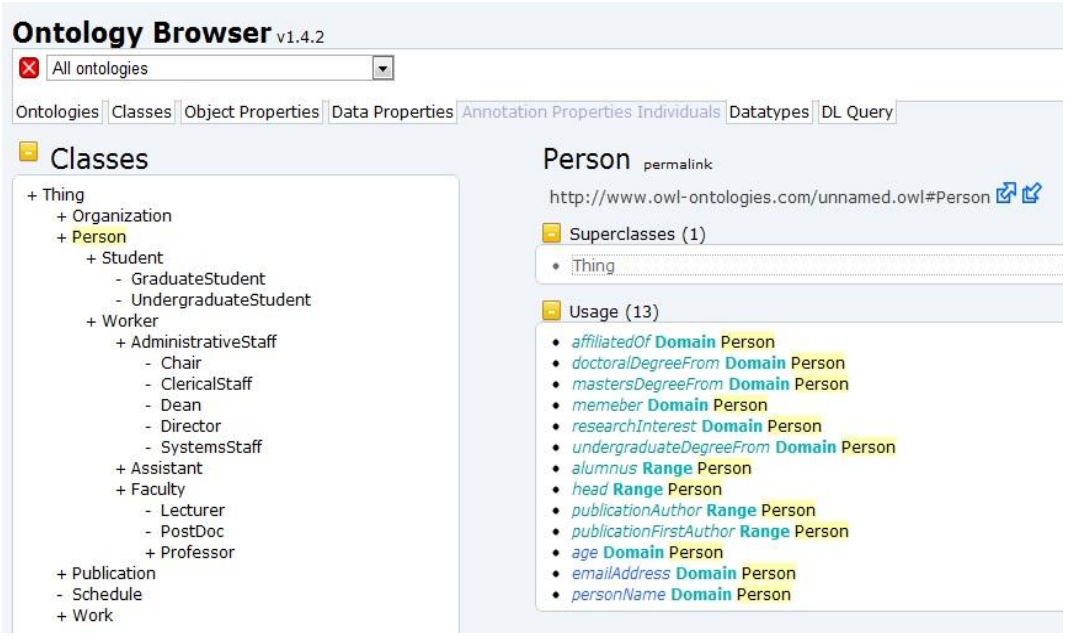


Figura 5.1 – Trecho da Ontologia de Domínio de (Yuan *et. al.* 2006)

O conceito **#Student** era especializado somente em **#Undergraduatestudent** e **#GraduateStudent** e também foi especializado em **#Bolsista** e **#Avulso.** O conceito **#Undergraduatestudent,** filho de **#Student**, foi especializado em #**Estagiario** e o conceito **#GraduateStudent,** filho de **#Student**, foi especializado em **#Mestrando** e **#Doutorando**. O conceito **#Usuario** foi criado como a união dos conceitos **#Student** e **#Professor** para descrever a regra definida para o sistema de controle da biblioteca que qualquer um destes pode ser cadastrado como usuário da biblioteca e realizar empréstimo de itens do acervo.

Todos os conceitos, atributos e relacionamentos que foram mantidos e os que foram incluídos receberam pelo menos um rótulo em português. Em alguns foi incluído mais de um rótulo como, por exemplo, o conceito **#Student** que foi associado aos termos ***"Aluno"*** e ***"Estudante"***, o conceito **#Professor** denominado pelos termos ***"Professor"*** e ***"Docente"*** e o conceito **#Instituicao_Federal_de_Ensino_Superior** que além do rótulo com o mesmo nome do conceito foi associada a sigla ***"IFES"***.

Além disto, algumas instâncias foram criadas na ontologia com o objetivo de exemplificar indivíduos que pertencem aos conceitos modelados, mas estas instâncias não foram utilizadas durante a atividade de busca, sendo visualizadas somente através da navegação na ontologia. A instância **#DIA,** com os rótulos ***"Departamento de Informática Aplicada"*** e ***"DIA"*** do conceito **#Department** é um exemplo.

5.1.2 Análise do Esquema de Metadados

O esquema de metadados adotado para o experimento foi proposto por Py *et. al.* (2009). Este esquema contém elementos que descrevem serviços dependentes de domínio, divididos em serviços de dados e serviços de conceito, para suportar a integração de informações em uma arquitetura orientada a serviços. Este esquema foi analisado para selecionar os elementos descritivos que fossem aplicáveis à busca por palavras-chaves, os elementos a serem utilizados para anotação semântica e os elementos que representam critérios de qualidade.

Uma vez que este esquema foi implementado no framework de Py *et. al.* (2009) como um banco de dados relacional, a análise dos elementos foi realizada em relação as tabelas e colunas do seu modelo físico. As colunas **conceito** e **descricao** da tabela **conceito**, as colunas **esquema** e **descricao** da tabela **esquema** e a coluna **descricao** das tabelas **servicoDados** e **servicoConceitos**, que são especializações da tabela **servico**, foram selecionadas para utilização na busca textual através da lista de palavras-chave semanticamente expandida. A regra de preenchimento definida no método proposto por Py *et. al.* (2009) para a coluna **nome** da tabela **servico** determina que, para serviços de dados, o nome do serviço seria a concatenação de uma parte fixa (srvDados) com o nome da fonte e o nome do esquema e, para serviços de conceitos, o nome do serviço seria a concatenação de uma parte fixa (srvConceito) com o nome do conceito.Em função disto esta coluna não foi utilizada pois a sua utilização iria requerer a remoção da parte fixa (srvDados e srvConceito), uma vez que esta não agrega semântica para fins de localização. Além disto, seria necessária a aplicação de um método de segmentação de palavras no conteúdo deste elemento cujo resultado seria redundante em relação ao conteúdo de outros elementos que foram selecionados para a busca.

Neste esquema de metadados existe somente um elemento associado à qualidade das fontes de dados. Este elemento representa a prioridade dos dados de um esquema em relação a outro, é aplicável quando um atributo do conceito é redundante e corresponde à coluna **prioridade** da tabela de relacionamento entre as tabelas **conceito** e **esquema** (**conceitoesquema**). Considerando esta particularidade do esquema, não foi necessária a seleção do critério de qualidade pelo participante e este elemento foi utilizado somente para ordenação dos resultados da busca.

O esquema de metadados foi alterado para adicionar um elemento responsável por armazenar as anotações em relação à ontologia de domínio. Uma coluna chamada **uri** foi acrescentada à tabela **conceito**. Esta tabela foi selecionada por armazenar os dados referentes aos conceitos de dados identificados nas

atividades de um processo organizacional, sendo um conceito de dados definido como uma entidade abstrata que representa um conjunto de entidades concretas, distribuídas em diferentes fontes e que compartilham a mesma semântica no escopo do negócio.

Uma vez que não existiam modelos de negócio e nem glossários para a aplicação do método, foi realizada a análise dos modelos das fontes de informação dos sistemas de informação descritos no item 5.1.3.

5.1.3 Sistemas de Informação

O cenário criado para o experimento, ilustrado na figura 5.2, caracterizou um ambiente com fontes de informação heterogêneas e autônomas, composto por cinco diferentes sistemas de informação que suportam as diversas atividades de uma instituição pública de ensino superior. Cada sistema é descrito a seguir e os respectivos modelos de dados se encontram no anexo II (Modelos de Dados do Experimento).

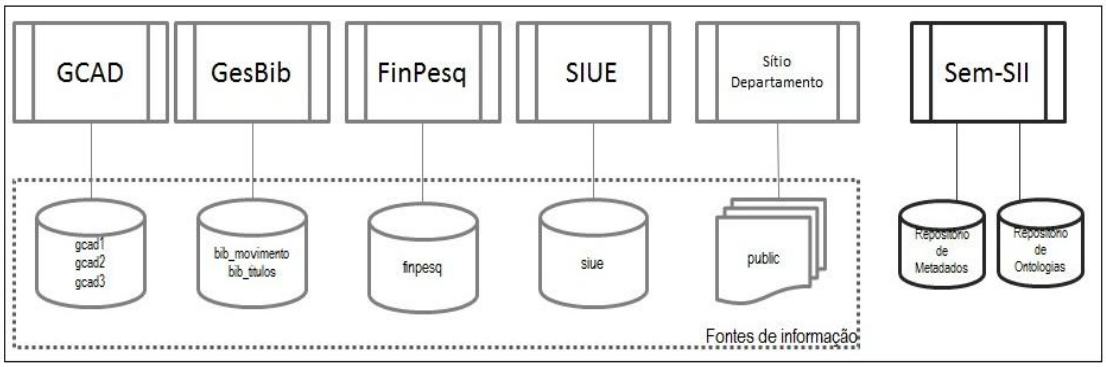


Figura 5.2: Cenário de Aplicação do Experimento

Gestão Acadêmica (GCAD)

As informações dos alunos, horários das aulas, professores de disciplinas, grade curricular dos cursos, matrícula e notas dos alunos em disciplinas e alocação de salas, auditórios e laboratórios para aulas estão registradas em um sistema que suporta as atividades de gestão acadêmica. Estas informações são atualizadas pelos funcionários administrativos da secretaria de cada departamento e pelos professores.

A base de dados se encontra armazenada em *PostgreSQL* 8.4 e foi separada em três esquemas: gcad1, gcad2 e gcad3 para permitir a definição das permissões de acesso às tabelas de acordo com o perfil dos usuários. Os respectivos modelos lógicos estão representados nas figuras A2.1, A2.2 e A2.3 do anexo II e a descrição na tabela 5.2.

Bib-online (GesBib)

Um terceiro sistema faz o controle da biblioteca. Neste sistema, estão cadastrados todos os títulos do acervo e o movimento de empréstimo, permuta e alienação de exemplares, caso sejam inutilizados. As informações são atualizadas somente pela bibliotecária através de um aplicativo instalado em seu desktop, mas o histórico de empréstimos e disponibilidade de itens do acervo podem ser consultados por todos os usuários via web. A base de dados se encontra armazenada em *PostgreSQL* 7.1 e foi dividida em dois esquemas: bib_titulos e bib_movimento (figuras A2.5 e A2.6 do Anexo II).

Financiamento em Pesquisa (FinPesq)

Em outro sistema, estão registrados dados das instituições financiadoras, dos projetos de pesquisa, dos recursos financeiros gastos com a participação de integrantes dos projetos em eventos como seminários, congressos e cursos, com a aquisição de material e equipamentos e com o pagamento de ajuda de custo aos bolsistas. Este sistema é atualizado pela comissão de concessão de bolsas e pelos coordenadores dos projetos de pesquisa e consultado pelas instituições financiadoras e pelos demais participantes dos projetos. A base de dados se encontra armazenada em Oracle 11g e o seu respectivo modelo é apresentado na figura A2.7 do Anexo II.

Sistema de Integração Universidade-Empresa (SIUE)

O último sistema é mantido pelos professores que realizam o acompanhamento de estágio com informações de alunos que realizam ou já realizaram atividade de estágio supervisionado, além de empresas com convênio com a Universidade e de vagas em aberto para estágio. O SGBD usado foi *Oracle* 10g e o seu respectivo modelo é apresentado na figura A2.8 do Anexo II.

Sítios dos Departamentos

O sítio de cada departamento possui uma página por disciplina com a sua descrição, ementa e o material de referência. Estas informações são mantidas pelos coordenadores do curso e pelos próprios professores através de um Sistema de Gerenciamento de Conteúdo (*Content Management System* – CMS). Os documentos estão armazenados em formato XML e seguem a especificação do esquema XML apresentado na figura A2.4 do anexo II.

A descrição das fontes, esquemas e conceitos de dados encontrados nestes sistemas foi realizada com o protótipo através da opção '**Registrar Metadados',** de acordo com as tabelas 5.1 e 5.2. Além disto, cada esquema foi associado ao seu respectivo modelo lógico, como apresentado nas figuras A2.1 a A2.8 do Anexo II, simulando uma integração entre o repositório de metadados do sistema de integração de informações e do repositório de modelos de dados institucional.

Tabela 5.1 Fontes, Esquemas e Serviços de dados

Fonte	**Esquema / Descrição**	**Serviço Dados / Descrição**
Gcad	gcad1 / Contém informações sobre professor e outros funcionários, departamento, cursos e suas disciplinas.	srvDadosGcad1 / Acesso aos dados do esquema gcad1 do sistema Gerencia Acadêmica.
Gcad	gcad 2 / As tabelas deste esquema armazenam info sobre alunos, matrículas, inscrições em disciplinas e suas notas e turmas.	srvDadosGcad2 / Acesso aos dados do esquema gcad2 do sistema Gerencia Acadêmica.
gcad	gcad3 / Este esquema é usado para a grade de horários de aulas e alocação de salas.	srvDadosGcad3 / Acesso aos dados do esquema gcad3 do sistema Gerencia Acadêmica.
Site departamento	public / As disciplinas de cada curso, com ementa, horário e bibliografia, são publicadas no site de cada departamento.	srvDadosSite / Obtém conteúdo de documentos xml do site.
bib-online	bib_movimento / Movimento de empréstimos e devoluções de exemplares feitos por usuários.	srvDadosBib-movimento / Acesso ao histórico de movimento (devolução e empréstimo) do acervo da biblioteca registrado no sistema GesBIB.
bib-online	bib_titulos / Dados sobre títulos, autores e exemplares disponíveis no acervo.	srvDadosBib-titulos / Acesso ao catálogo do acervo da biblioteca que estão no sistema GesBIB.
FINPESQ	finpesq / Controle de projetos de pesquisa, principalmente no que diz respeito a custos.	srvDadosFinPesq / Acesso aos dados do sistema FINanciamento de PESQuisa.
SIUE	siue / Contém informações sobre ofertas de vagas e acompanhamento de atividades realizadas pelos alunos nas empresas.	srvDadosSIUE / Acesso aos dados do Sistema de Integração Universidade-Empresa (SIUE)

Tabela 5.2 Conceitos, Serviços, Prioridade dos Esquemas e Anotação Semântica

Conceito de Dados	**Serviço Conceito**	**Esquema / Prioridade**	**Anotação Semântica**
aluno	srvConceitoAluno	gcad2 / 1	...#Student
aluno de graduação	srvConceitoAlunoGraduacao	gcad2 / 1	...#UndergraduteStudent
aluno de mestrado	srvConceitoAlunoMestrado	gcad2 / 1	...#Mestrando
Aquisições	srvConceitoAquisicao	finpesq / 1	
Bolsa	srvConceitoBolsa	finpesq / 1	...#Auxilio_Financeiro
Cursos	srvConceitoCursos	gcad1 / 1 public / 2	...#Course
Cursos externos	srvConceitoCursosExternos	siue / 1	
departamentos	srvConceitoDepartamentos	gcad1 / 1	...#Department
devolução	srvConceitoDevolucao	bib_movimento / 1	
Disciplinas	srvConceitoDisciplinas	gcad1 / 1 public / 2	
Empresa	srvConceitoEmpresa	siue / 1	...#Empresa

empréstimo	srvConceitoEmprestimo	bib_movimento / 1	...#Empresta
Equipamento	srvConceitoEquipamento	finpesq / 1	
Equipe de projeto	srvConceitoEquipeProjeto	finpesq / 1	
Estágio	srvConceitoEstagio	siue / 1	...#Estagiario
exemplar	srvConceitoExemplar	bib_titulos / 1	
financiadora	srvConceitoFinanciadora	finpesq / 1	
Laboratório	srvConceitoLaboratorio	gcad3 / 1	
Livro	srvConceitoLivro	bib_titulos / 1 public / 2	...#Book
Localização Aulas	srvConceitoLocalizacaoAulas	gcad3 / 1	
Participação em Eventos	srvConceitoParticipacaoEventos	finpesq / 1	
Perfil Candidato	srvConceitoPerfilCandidato	siue / 1	...#UndergraduteStudent
Produto Final de Curso	srvConceitoProdutoFinalCurso	bib_titulos / 1	...#Trabalho_Teorico
professor	srvConceitoProfessor	gcad1 / 1	...#Professor
Programa	srvConceitoPrograma	public / 1	...#Software
projeto de pesquisa	srvConceitoProjetoPesquisa	finpesq / 1	
Sala de Aula	srvConceitoSalaAula	gcad3 / 1	
Secretaria	srvConceitoSecretaria	gcad1 / 1	...#Secretaria
Supervisior de Estágio - acadêmico	srvConceitoSupervisiorEstagioAcademico	siue / 1	...#Orientador
Supervisior de Estágio - profissional	srvConceitoSupervisiorEstagioProfissional	siue / 1	
títulos	srvConceitoTitulos	bib_titulos / 1	...#Item_Acervo
turma	srvConceitoTurma	gcad1 / 1 gcad1 / 2	
Usuário	srvConceitoUsuario	bib_movimento / 1	
Vaga Estagio	srvConceitoVagaEstagio	siue / 1	

5.1.4 Dinâmica do Experimento

A partir da descrição dos sistemas, seus modelos e metadados, foram elaborados três cenários com diferentes necessidades de informações, conforme tabela 5.3. Para cada necessidade foram elencados os pares **ConceitoDeDados-Esquema** relevantes que continham as informações necessárias para atender a cada uma. Cada par **ConceitoDeDados-Esquema** representa uma fonte de informação e equivale a um documento em sistemas de recuperação de informações tradicionais.

Tabela 5.3 Necessidades de informação exploradas no experimento

1. O coordenador de um curso de graduação precisa saber se a previsão de conclusão do curso para os bolsistas que estão no 7º período está dentro do tempo de concessão da bolsa. Para aqueles que a previsão de conclusão seja posterior ao fim da bolsa, será necessário iniciar um processo de prorrogação junto à instituição financiadora.	
Esquema	**Conceito de Dados**
Finpesq	Bolsa

finpesq	Financiadora
gcad1	Cursos
gcad1	Disciplinas
gcad2	Aluno de graduação
2. A secretária de cada departamento precisa entregar no final de cada semestre uma relação de livros para serem comprados e que ficarão disponíveis para empréstimo aos alunos na biblioteca. A relação deve ser elaborada com base na bibliografia indicada pelas disciplinas que serão oferecidas no próximo semestre.	
Esquema	**Conceito de Dados**
bib_titulos	Livro
Public	Livro
gcad1	Cursos
gcad1	Disciplinas
public	Disciplinas
gcad1	Secretaria
gcad2	Turma
3. A secretária do departamento X está com os diplomas que acabaram de chegar do DRCA (Departamento de Registro e Controle Acadêmico). O diretor do departamento orientou que somente os alunos sem pendências na biblioteca poderão retirar o diploma. Aquele que tiver algum exemplar pendente de devolução ou que não entregou uma cópia impressa da monografia/dissertação/tese não poderá retirar o diploma. O coordenador da comissão de bolsa orientou que, em caso de ex-bolsistas, seja confirmado se não existe pendência em relação ao relatório semestral de atividade ou se a versão eletrônica da dissertação ou tese foi entregue. O aluno que tiver com pendência deverá procurar o coordenador do projeto de pesquisa para regularizar a sua situação, isto não impedirá a retirada do diploma. Somente após retirar o diploma, os alunos são considerados egressos.	
Esquema	**Conceito de Dados**
bib_movimento	empréstimo
bib_titulos	exemplar
bib_movimento	Usuário
bib_titulos	Produto Final de Curso
gcad2	Aluno
Finpesq	Bolsa
Finpesq	projeto de pesquisa
Finpesq	Equipe de projeto

Um roteiro para utilização do protótipo (Anexo III) foi distribuído junto com o convite para participação no experimento e neste documento constava uma breve descrição do escopo de cada sistema e o passo a passo para utilização do protótipo. Mas os participantes não foram apresentados previamente ao modelo de dados destes sistemas, ao esquema de metadados usado para catalogá-los, à ontologia de domínio

que foi utilizada para anotação e nem aos metadados que descrevem os conceitos de dados existentes nestes sistemas.

O protótipo apresentou aos usuários a descrição das três necessidades de informação e os participantes foram orientados a realizar consultas no repositório de metadados, através da opção '**Buscar**', para localizar as fontes de informação que atendem a cada necessidade apresentada. Para cada fonte de informação recuperada são apresentados os metadados, o modelo de dados e o conceito da ontologia, quando tiver sido anotada. A aplicação permitiu que o participante pudesse explorar tanto o modelo quanto à anotação para avaliar se a fonte de informação era relevante para o cenário em questão. No decorrer do experimento, os participantes também puderam explorar a ontologia através da opção '**Visualizar Ontologia'** usando o *ontology browser*.

5.2 Análise Quantitativa e Qualitativa do Experimento

A pesquisa quantitativa requer a definição de variáveis observadas, que são objetivas e medidas em escalas numéricas, e tem por objetivo verificar o quanto uma abordagem proposta é melhor em relação às demais abordagens possíveis para solucionar o mesmo problema. As variáveis a serem observadas são consideradas objetivas se diferentes observadores obtêm os mesmos resultados em observações distintas e existe consenso no que diz respeito aos valores esperados dessas variáveis (Wainer 2007).

A análise quantitativa das variáveis dependentes (precisão, cobertura e medida F) determina o desempenho de cada abordagem a partir da manipulação de variáveis independentes (número de consultas executadas, termos informados pelo usuário, recursos selecionados pelo usuário). Assim o experimento teve a seguinte configuração:

Variáveis de controle

- Quantidade de recursos da ontologia: 159
    - Quantidade de Classes/Conceitos: 92
    - Quantidade de Propriedades de objetos/Relacionamentos: 39
    - Quantidade de Propriedades de dados/Atributos: 28
- Quantidade de Conceitos de Dados do repositório: 34
    - Quantidade de Conceitos de Dados anotados: 16
- Sistemas de Informação: 5
- Necessidades de informação: 3

- Quantidade de fontes de informação relevantes em cada necessidade de informação

Variáveis Independentes

- Termos iniciais das consultas
- Recursos selecionados pelo usuário para expansão

Variáveis Dependentes de cada abordagem usada na comparação

- Cobertura
- Precisão
- Medida F1

5.2.1 Abordagens para Comparação dos Resultados da Busca

Quatro abordagens foram utilizadas para comparação dos resultados da busca:

1. **Consulta original**: busca textual por fontes de informação usando a lista de palavras-chave inicial de cada consulta.
2. **Expansão semântica automática**: busca semântica e busca textual com a lista de palavras-chave semanticamente expandida utilizando todos os rótulos e URIrefs dos recursos sugeridos pelo sistema.
3. **Expansão estatística automática**: busca textual com a lista de palavras-chave estatisticamente expandida utilizando padrões de co-ocorrência de termos dos elementos descritivos dos metadados.
4. **Expansão semântica assistida pelo usuário**: busca nas anotações e busca textual com a lista de palavras-chave semanticamente expandida seguindo a arquitetura proposta.

A abordagem da consulta original representa o *baseline* de comparação das demais abordagens. Em todas as abordagens a lista de palavras-chaves inicial de cada consulta foi submetida à lematização, *stemming* e remoção de *stopwords.*

As duas abordagens de expansão semântica (automática e assistida pelo usuário) utilizam a ontologia de domínio como um modelo de conhecimento independente do corpus para expansão da consulta (Bhogal *et. al.* 2007). A expansão semântica automática simulou uma alternativa à arquitetura proposta na qual a expansão da consulta não requer a intervenção do usuário para selecionar os recursos da ontologia sugeridos pelo sistema.

A abordagem de expansão estatística automática é baseada em um modelo de conhecimento dependente do corpus de documentos (Bhogal *et. al.* 2007), neste caso, o repositório de metadados. A criação deste modelo demandou a geração de duas tabelas de apoio para extrair padrões de co-ocorrência de termos chamadas

termoConsulta-Metadados e **termoConsulta-termoMetadados-freqTotal**. Os elementos selecionados do esquema de metadados foram submetidos à lematização, *stemming* e remoção de *stopwords* antes da geração das tabelas e a geração destas tabelas requereu a realização dos seguintes passos:

1. Para cada termo distinto informado pelo usuário (**termoConsulta**) foram identificados os metadados das fontes de informação onde este termo ocorre e esta combinação foi registrada em uma tabela (**termoConsulta-Metadados)**.
2. A partir de cada fonte de informação, foram identificados todos os termos distintos (**termoMetadados**) e suas respectivas freqüências dentro dos metadados desta fonte (**freqMetadados**). A freqüência foi calculada usando a função *ts_stat* do *Full Text Search*.
3. A tabela **termoMetadados-termoMetadados-freqTotal** foi atualizada considerando que: (a) se a combinação **termoConsulta** e **termoMetadados** não existe, então a combinação deve ser registrada com a freqüência de ocorrência da primeira fonte de informação identificada e (b) se já existir então a freqüência de ocorrência da nova fonte de informação deve ser acumulada na freqüência que se encontrava registrada.

Para cada consulta registrada no *log,* foi criada uma nova lista de palavras-chave adicionando os termos (**termoMetadados**) com maior freqüência de co-ocorrência (**freqTotal**) em relação aos termos da lista inicial (**termoConsulta**) a partir da tabela resultante **termoConsulta-termoMetadados-freqTotal**. O cálculo das medidas de desempenho da abordagem de expansão estatística automática utilizou o resultado desta nova lista de palavas-chave gerada.

5.2.2 Resultados do Experimento

Trinta e sete pessoas, entre alunos de mestrado da UNIRIO, pesquisadores do NP^{2}TEC e analistas de sistemas do IBGE, participaram do experimento. No total foram realizadas 141 consultas distintas durante o período previsto para coleta de dados. Do conjunto total de consultas formuladas pelos participantes a partir da descrição de cada necessidade de informação, nove consultas não recuperaram nenhum conceito de dados e foram descartadas. A distribuição de consultas por necessidade de informação apresentada foi a seguinte: 51 consultas para a primeira, 45 consultas na segunda e 36 consultas na terceira. A eficiência do resultado de cada abordagem foi

avaliada usando a média da precisão, cobertura e da medida F1 das 132 consultas restantes, conforme apresentado no gráfico da figura 5.3. A medida F1, é o cálculo da medida F onde a precisão e a cobertura têm peso igual a 1.

A abordagem de expansão semântica assistida pelo usuário, que segue a arquitetura proposta, atingiu a maior precisão na média e este resultado favorece os cenários onde a quantidade de fontes de informação catalogadas no repositório de metadados apresenta tendência de crescimento. A quantidade de falsos positivos recuperados foi menor que as abordagens de expansão semântica automática e expansão estatística. A justificativa para este resultado é que o controle dos usuários sobre as expansões realizadas permitem que este esclareça o contexto de sua intenção de busca e evite a perda do foco da consulta.

A comparação entre o resultado da busca com a consulta original e as buscas através da expansão semântica (automática ou assistida pelo usuário) demonstra que tanto a precisão quanto a cobertura obtiveram melhores resultados em função da ontologia de domínio restringir a expansão aos termos do domínio. Além disto, a existência de conceitos de dados anotados em relação à ontologia permitiu a recuperação através dos ponteiros dos conceitos da ontologia, mesmo quando os termos informados pelo usuário não estavam presentes no conteúdo dos metadados descritivos.

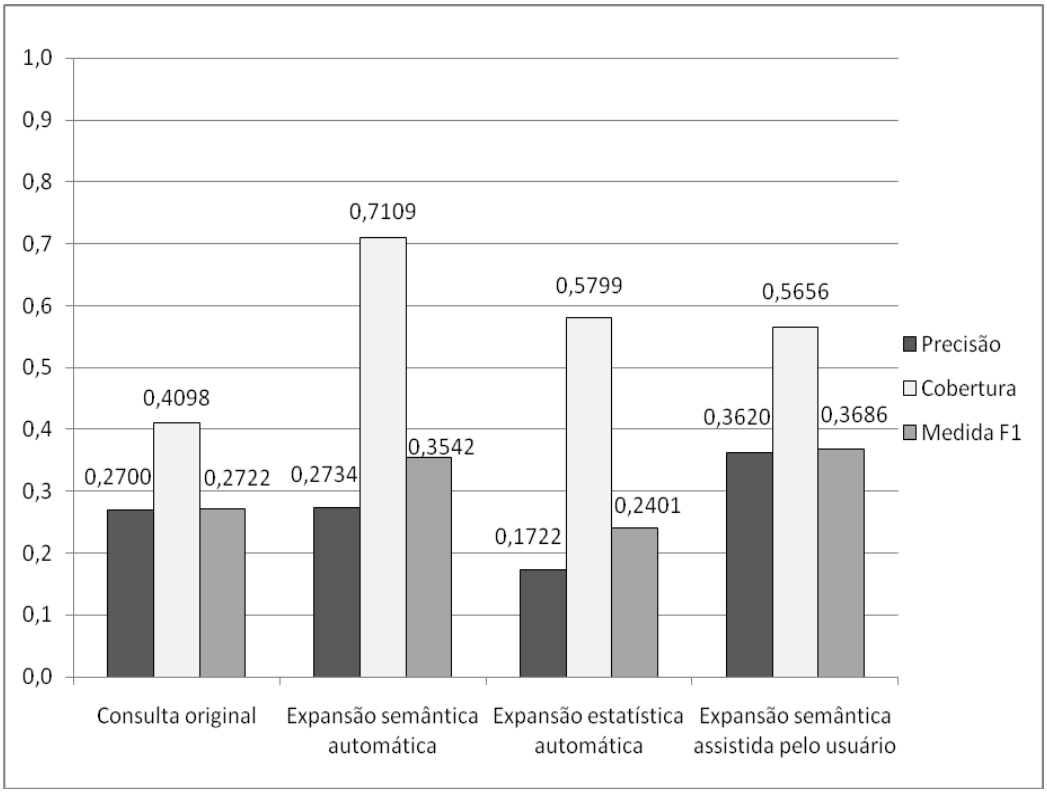


Figura 5.3 – Média da Precisão, Cobertura e Medida F1

A média da cobertura do resultado das buscas através da expansão semântica automática foi maior que a assistida pelo usuário enquanto que a média da precisão da primeira foi menor que a segunda. Este resultado era esperado uma vez que o conjunto de recursos utilizados para expansão assistida pelo usuário é um subconjunto dos recomendados pelo sistema. Porém, além da expansão assistida pelo usuário recuperar menos falsos positivos e requerer que o usuário analise um número menor de metadados, o usuário tem a oportunidade de adquirir mais conhecimento sobre o domínio e o vocabulário utilizado à medida que ele avalia os fragmentos da ontologia recuperados pelo sistema como recomendações para a expansão.

O cálculo da precisão, cobertura e medida F1 para a expansão estatística automática considerou o resultado da busca no repositório de metadados com a nova consulta gerada a partir das tabelas de co-ocorrência. A cobertura (0,5799) obteve um valor um pouco maior que a expansão semântica assistida pelo usuário (0,5656), mas a precisão da primeira (0,1722) foi menos que a metade da segunda (0,3620) e também menor que a abordagem da consulta original.

Em todos os cenários de comparação, a abordagem de expansão semântica assistida pelo usuário obteve melhores resultados na medida F1, refletindo o equilíbrio entre o aumento da precisão e da cobertura obtido com a arquitetura. A medida F1 é a média harmônica ponderada entre precisão e cobertura atribuindo a ambas o mesmo peso igual a 1.

A análise do *log* de consultas de ferramentas de busca permite obter outras informações estatísticas como, por exemplo, a quantidade de termos por consulta e o número de consultas por usuário além de proporcionar a identificação de padrões de modificação de consultas por parte dos usuários. A modificação da consulta pode ser analisada através de pares de consultas submetidas sucessivamente em uma mesma sessão de busca. Na maioria dos estudos avaliados em (Hollink *et. al.* 2011), as modificações são classificadas em especificação, quando termos são adicionados, generalização, se algum termo for removido e reformulação quando pelo menos um termo é substituído por outro. Neste artigo também foram citados alguns estudos, que avaliam a semântica dos termos usados na consulta, a substituição de um termo por outro pode significar a especialização da consulta se o termo substituto representar um hipônimo do termo inicial. Outros estudos classificam a especificação, generalização e reformulação da consulta em função da freqüência de ocorrências de seus termos na coleção de documentos usada. Nenhuma consideração é feita quanto ao papel dos operadores (E e OU) na modificação das consultas.

Neste experimento, a avaliação manual do *log* revelou que em algumas buscas os usuários repetiram a consulta reduzindo ou adicionando palavras, mudando o operador de E (&) para OU (|) e mudando o conjunto de recursos selecionados para expansão. A tabela 5.4 exemplifica estes casos de acordo com o tipo de modificação realizada. A especificação foi considerada como a remoção de termos em uma consulta com o operador OU (|), pois resulta na redução do número de fontes de informação recuperadas e a generalização como a inclusão de termos com o operador OU (|). A reformulação foi identificada quando o usuário repetiu a mesma consulta com o mesmo operador, mas diferentes recursos foram selecionados para expansão. Este comportamento pode resultar em redução ou aumento do número de fontes de informação recuperadas. Uma quarta classificação foi criada para caracterizar as situações onde houve somente a mudança no operador para combinação dos termos de E (&) para OU (|), resultando sempre em aumento do número de fontes de informação recuperadas.

A lista de palavras-chave no protótipo estava limitada, por questões técnicas da linguagem, em seis termos por consulta e a média de termos informados por consulta foi de 4,39 no universo total de 141 consultas. A tabela 5.5 demonstra a distribuição das consultas em relação aos termos informados.

Tabela 5.4 Exemplos de Modificação de Consultas

Modificação da consulta pelo usuário	Lista inicial de palavras-chave	Fontes de informação recuperadas	Necessidade de informação
Especificação	diploma | aluno | pendência | biblioteca | bolsista	19	3
	pendência | biblioteca	3	3
Generalização	Compras	1	2
	compras | biblioteca	4	2
Reformulação	departamento | semestre | livros | biblioteca | bibliografia | disciplinas	25	2
	departamento | semestre | livros | biblioteca | bibliografia | disciplinas	17	2
Flexibilização	departamento & aluno & livro & data & disciplina & semestre	3	2
	departamento | aluno | livro | data | disciplina | semestre	25	2

Tabela 5.5 Distribuição de Consultas *versus* Quantidade de Termos

Quantidade de termos (X)	Quantidade de consultas com X termos encontrados		
	na lista de palavras-chave	na ontologia	nos metadados
0	*	9	6
1	15	21	15
2	11	31	23
3	15	**33**	**35**
4	21	21	24
5	33	16	22
6	**46**	10	16

Dentre os setenta e três termos distintos informados nas consultas, trinta e nove termos não foram encontrados na ontologia, trinta e três não foram encontrados nos metadados descritivos e vinte e seis não foram encontrados em ambos. Removendo alguns termos que estavam fora do domínio e com erros ortográficos, os termos restantes foram analisados e utilizados na evolução da ontologia para adição de rótulos a conceitos existentes e de recursos (conceitos, relacionamentos e atributos). Alguns exemplos podem ser citados como a adição dos conceitos **#Diploma**, **#Egresso** e **#Reprovacao**, com rótulos do mesmo nome, e de rótulos como ***"ex-Aluno"*** para o conceito **#Egresso,** ***"Matéria"*** para o conceito existente **#Disciplina,** ***"TCC"*** para o conceito **#Monografia** e ***"Referências"*** para o conceito **#Material_de_Aprendizado**. Vale ressaltar que as palavras ***"Matéria"*** e ***"Egresso"*** são termos polissêmicos, mas que podem ser interpretados como sinônimos para as palavras ***"Disciplinas"*** e ***"ex-Aluno"*** dentro deste domínio.

Após a adição do termo ***"Matéria"*** como rótulo do conceito **#Disciplina** foi solicitado ao usuário que utilizou este termo em sua busca que repetisse esta consulta, pois na primeira vez nenhum recurso havia sido encontrado na ontologia e nenhuma fonte de dados nos metadados. Desta vez o sistema recuperou o conceito **#Disciplina** e dezessete outros recursos associados através da busca com propagação. O usuário selecionou o conceito **#Disciplina** e o conceito relacionado **#Course** e vinte e uma fontes de informação foram recuperadas do repositório de metadados. Deste resultado três eram relevantes, uma foi recuperada usando o ponteiro do conceito **#Course** nas anotações e as duas outras através dos termos ***"Disciplina"*** e ***"Curso",*** adicionados à lista de palavras-chave para a busca textual.

Apesar dos usuários terem sidos orientados para informar somente termos que representam conceitos na busca, a análise do *log* demonstrou que foram informados termos que representam recursos que foram modelados como atributos e relacionamentos na ontologia. Esta situação sugere que a perspectiva de quem realiza a busca quanto ao que vem a ser um conceito, relacionamento, atributo ou instância nem sempre corresponde a de quem modelou a ontologia. O termo **"*Autor"*** foi usado cinco vezes por usuários diferentes e na ontologia este está associado com rótulo de um atributo do conceito **#Item_Acervo**, outro exemplo deste mesmo tipo é o termo ***"ISBN"*** que corresponde a um atributo do conceito **#Book**.

Foram encontrados no *log* termos que correspondem a relacionamentos que ainda não estavam modelados na ontologia como, por exemplo, o termo ***"requisito"*** que nomeia o relacionamento entre disciplinas na forma **#Disciplina #eh_Pre_Requisito_De #Disciplina**. Os participantes também utilizaram a sigla

"***DRCA***" em seis consultas associadas à terceira necessidade de informação apresentada. Esta sigla se refere ao nome do departamento que emite os diplomas *"Departamento de Registro e Controle Acadêmico"* e constava na descrição desta necessidade de informação somente para tornar a apresentação da mesma mais completa. Na ontologia foi incluída uma instância do conceito **#Department** especificando a sigla e o nome desta unidade organizacional de modo a esclarecer que não se trata de um conceito.

Os termos ***"Bolsa"***, usado em 26 consultas, e ***"Bolsista"***, que ocorreu em 24 consultas, possuem o mesmo radical ***"Bolsa".*** De acordo com o dicionário *Michaelis* (Weiszflog 2006) a palavra ***"Bolsa"*** possui mais de um significado (polissêmica), mas o único significado aplicável a este contexto é *"**Bolsa de Estudos:** soma de dinheiro ou seu equivalente, oferecida por uma instituição educacional ou outra qualquer entidade pública ou particular, para possibilitar ou facilitar a um estudante o prosseguimento de seus estudos em uma escola ou universidade".* Por outro lado, o adjetivo ***"Bolsista"*** é formado por sufixação (bolsa+ista) e este conceito classifica alunos como *"Quem recebeu bolsa de estudos"*.

Nos casos onde as palavras-chave informadas pertencem ao mesmo campo lexical, o resultado da busca com qualquer um dos termos na ontologia recupera o mesmo conjunto de recursos. Dentro deste conjunto estão o conceito **#Auxilio_Financeiro**, pois um de seus rótulos é o termo ***"Bolsa"***, e o conceito **#Bolsista**, com o rótulo de mesmo nome e conceito filho de **#Student**. Uma vez que o sistema apresenta este conjunto de recursos ao usuário, é possível que este esclareça qual conceito deseja. Quando o usuário utiliza o termo ***"Bolsa"*** e a sua intenção de busca está associada ao conceito **#Auxilio_Financeiro**, o ponteiro deste conceito da ontologia usado para consultar as anotações torna o resultado mais preciso ao recuperar o conceito de dados Bolsa. Porém, se a intenção de busca estiver associada ao conceito **#Bolsista**, a adição deste termo não melhora o resultado da busca textual nos elementos descritivos, se comparado com a busca com a lista original de termos, pois esta característica do radical em comum também ocorre nesta busca. O conjunto de metadados recuperado neste caso é o mesmo se nenhum outro recurso sugerido pelo sistema for selecionado pelo usuário, além disto, não existia um conceito de dados chamado Bolsista nos metadados.

O termo ***"Biblioteca"*** foi utilizado em trinta consultas diferentes que pertencem ao conjunto de consultas da segunda e terceira necessidade de informação apresentada. O efeito da expansão usando este termo é apresentado na tabela 5.6. O exemplo de expansão semântica assistida pelo usuário apresentado na tabela foi extraído do *log* de um usuário em particular e foi escolhido em função da precisão e

cobertura do resultado em relação à abordagem de expansão semântica automática. Neste caso, a segunda opção recuperou as mesmas três fontes de informação relevantes que a terceira porém o total de fontes de informação retornados na segunda foi metade. A quarta opção, expansão estatística automática, teve duas fontes de informação adicionais recuperadas em relação às duas opções anteriores ao custo de um total de 19 fontes de informação recuperadas.

Tabela 5.6 Efeito das Abordagens de Expansão na Lista de Palavras-chave

Abordagem	Lista de palavras-chave	Total de fontes de informação recuperadas	Fontes de informação relevantes recuperadas
Consulta original	biblioteca	3	2
Expansão semântica assistida pelo usuário	('biblioteca') | ('emprestar') | ('item' & 'acervo')	5	3
Expansão semântica automática	('biblioteca') | ('emprestar') | ('item' & 'acervo') | ('vincular') | ('unidade' & 'acadêm') | ('pertence' & 'acervo')	10	3
Expansão estatística automática	('biblioteca') | ('exemplar') | ('mover') | ('movimentar') | ('movimento') | ('título')	19	5

Além de acrescentar recursos e rótulos na ontologia, os termos utilizados nas consultas dos participantes foram usados para anotar Conceitos de Dados existentes, como por exemplo Disciplinas (#Disciplina), e cadastrar novos Conceitos de Dados, já com anotação, como por exemplo Devolução (#Devolve). A tabela 5.8 apresenta estes e outros casos de anotação semântica e criação de Conceitos de Dados (em negrito).

Tabela 5.7 Fontes de Informação Anotadas e Acrescentadas após a Análise do *log* de Consultas

Conceito de Dados	Serviço Conceito	Esquema / Prioridade	Anotação Semântica
devolução	srvConceitoDevolucao	bib_movimento / 1	**…#Devolve**
Disciplinas	srvConceitoDisciplinas	gcad1 / 1 public / 2	**…#Disciplina**
Bolsista	srvConceitoBolsista	finpesq / 1 gcad1 / 2	...#Bolsista
Egresso	srvConceitoEgresso	gcad1 / 1	...#Egresso
financiadora	srvConceitoFinanciadora	finpesq / 1	**…#Apoio_a_Pesquisa**

5.3 Questionário aos Participantes

Ao final do experimento, os participantes receberam um questionário, subdividido em três seções com 16 perguntas no total. Do total de trinta e sete participantes, somente vinte responderam ao questionário. O primeiro grupo de perguntas chamado *"Sobre o participante"* composto por cinco perguntas teve por objetivo caracterizar o perfil do participante, sua experiência prévia e conhecimento sobre a busca por fontes de informações, conforme apresentado na tabela 5.8.

Tabela 5.8 Perguntas e Respostas da Seção "*Sobre o Participante*"

1) Você já trabalhou com modelagem de fontes de informações e de sistemas de informações?	
Sim, tenho mais de 10 anos de experiência.	9
Sim, entre 5 e 10 anos de experiência.	4
Não, meu conhecimento é somente teórico.	3
Não, mas já tive contato com modelos deste tipo.	2
Sim, até 5 anos de experiência.	2
2) Como você descreve o seu conhecimento sobre ontologia?	
Tenho conhecimento teórico e estou aprofundando este conhecimento dentro da minha pesquisa.	6
Nenhum, o meu primeiro contato foi através deste experimento.	5
Tenho conhecimento teórico mas não está vinculado a minha pesquisa.	3
Tenho conhecimento somente teórico.	3
Tenho conhecimento teórico e prático sobre ontologias.	2
Tive contato por alto no mestrado e depois num mini curso de Simpósio.	1
3) Você já utilizou um modelo de dados* para entender os conceitos existentes em uma fonte de informação? ** Modelos de dados são artefatos gráficos que podem usar notação ER, UML, ORM, etc*	
Sim, mas mesmo assim tive que recorrer a outras fontes como programas e colegas mais experientes para tirar algumas dúvidas.	10
Sim, e o modelo estava atualizado e suficientemente descrito para permitir o meu entendimento.	8
Não, nunca necessitei realizar este tipo de tarefa.	1
Apenas na faculdade.	1
Não, pois o modelo não existia ou não foi encontrado e por isso dependi exclusivamente de outras fontes como programas e colegas mais experientes.	0
4) Você já utilizou a descrição de fontes de informação (metadados*) para entender quais informações estavam presentes nestas fontes? ** De acordo com a NISO metadados são informações estruturadas que descrevem, explicam, localizam, ou tornam mais fácil a recuperação, utilização e gerenciamento dos recursos de informação.*	
Sim, mas mesmo assim tive que recorrer aos responsáveis pelas mesmas para tirar algumas dúvidas.	11
Sim, e os metadados estavam atualizados e claramente descritos para permitir o meu entendimento.	5
Não, pois os metadados não existiam ou não foram encontrados e por isso tive que depender exclusivamente dos responsáveis pelas mesmas para entendê-las.	3
Não, nunca necessitei realizar este tipo de tarefa.	1
5) Você já recorreu a dicionários especializados, glossários ou manuais técnicos para entender a descrição destas fontes de informação?	
Sim, isto já aconteceu algumas vezes.	10
Sim, isto já aconteceu várias vezes.	7
Não, quando necessário solicito (ou solicitarei) esclarecimentos adicionais ao responsável pela fonte de informação.	3
Não, tenho sólido conhecimento da terminologia utilizada.	0

A maioria dos respondentes tinha algum conhecimento sobre ontologias que variavam de teórico a prático, mas 25% deles tiveram contato com este artefato pela

primeira vez através do experimento. A média da precisão e da cobertura das consultas realizadas por este grupo de usuários foi calculada para comparação com a média de todos os participantes do experimento, conforme a tabela 5.9. O aumento da precisão deste grupo em particular foi bem menor que do grupo de todos os participantes enquanto que o aumento da cobertura ultrapassou o dobro da média geral.

Tabela 5.9 Média da precisão e cobertura – Comparativo do grupo com a média geral

		Consulta Original	Expansão semântica assistida pelo usuário	% de aumento
Média do grupo	Precisão	0,3378	0,3615	7,02 %
	Cobertura	0,3255	0,6068	86,42 %
Média Geral (Todos os Participantes)	Precisão	0,2700	0,3620	34,07 %
	Cobertura	0,4898	0,5656	15,48 %

A metade dos respondentes informou que já havia feito uso de modelos de dados e metadados para entender o conteúdo das fontes de informação e que também foi necessário recorrer a outras fontes para esclarecer suas dúvidas. Estas respostas revelam que mesmo quando os metadados e modelos das fontes de informação estão disponíveis para análise, estes ainda não revelam todos os aspectos semânticos das fontes de informação que descrevem e modelam. Neste cenário, a arquitetura apresentada permite o enriquecimento do conhecimento do usuário sobre o domínio, através da navegação na ontologia, e o aumento da precisão da descrição destas fontes, através das anotações.

Informações sobre o ambiente organizacional dos respondentes foram obtidas através do segundo grupo composto por oito perguntas, expostas através da tabela 5.10.

Tabela 5.10 Perguntas e Respostas da Seção "*Sobre o Ambiente Organizacional*"

6) Qual o ramo de atuação da organização que você está vinculado?	
Administração Pública	9
Petróleo e Gás	3
Financeiro	2
Educação e Pesquisa	2
Telecomunicações	1
Strategic Outsourcing em TI.	1
Mineração	1
Pesquisas e Estatísticas	1
7) Você já realizou alguma tarefa que envolvia <u>a localização</u> de fontes de informação nesta empresa para atender a uma necessidade de informação de um usuário de negócio?	
Sim, mas tive que entrar em contato com usuários e analistas pois as fontes de informação não estão registradas em repositórios.	8
Sim, mas tive que realizar diversas buscas, pois as fontes de informação estão registradas em vários repositórios.	5

Não, nunca necessitei realizar este tipo de tarefa e também nunca fui consultado a respeito.	5
Não, mas já fui consultado sobre a localização das fontes de informação dos sistemas que utilizo ou sou responsável.	2
Sim, e a busca foi realizada em um único repositório onde as fontes de informação estão registradas.	0
8) Você já realizou alguma tarefa que envolvia <u>o acesso</u> a mais de uma fonte de informação nesta organização para atender a uma necessidade de informação de um usuário de negócio?	
Sim, mas foi necessário entender a estrutura de cada uma das fontes de informação e realizar o acesso separadamente.	13
Não, nunca necessitei realizar este tipo de tarefa.	4
Não, mas já fui consultado sobre como acessar as fontes de informação dos sistemas que utilizo ou sou responsável.	2
Sim, e o acesso foi realizado através de um sistema de integração de informações.	1
Sim, realizei o acesso as fontes de informação separadamente através de uma interface padrão.	0
9) Nesta organização existem procedimentos definidos para modelagem e registro de novas fontes de informação?	
Não, mas algumas equipes desenvolveram alguns procedimentos próprios.	7
Sim, mas estes procedimentos são seguidos por somente algumas equipes pois não são obrigatórios.	4
Sim, mas estes procedimento são seguidos por somente algumas equipes pois não são amplamente divulgados.	3
Não sei / Desconheço	2
Sim, estes procedimentos são seguidos por todas as equipes pois são acompanhados por uma área de controle	2
Sim, mas estes procedimentos são seguidos por somente algumas equipes pois não são adequados as metodologias de desenvolvimento de sistemas.	1
10) Na sua organização existem normas que descrevem padrões de nomenclatura, domínio e tipos de dados?	
Sim, estas normas são seguidas por todas as equipes pois são pré requisitos para a implantação dos sistemas em produção.	5
Não, mas algumas equipes desenvolveram seus próprios padrões de nomenclatura, domínio e tipos de dados.	4
Sim, mas estas normas são seguidas por somente algumas equipes pois não são obrigatórios.	4
Sim, mas estas normas são seguidas por somente algumas equipes pois não são amplamente divulgadas.	3
Sim, mas estas normas são seguidas por somente algumas equipes que usam ferramentas de modelagem que permitem a definição destes padrões.	3
Não, as equipes não se preocupam padrões de nomenclatura, domínio e tipos de dados.	1
11) Na sua organização a atividade de administração de dados é realizada total ou parcialmente por:	
Uma equipe centralizada, dedicada exclusivamente para esta atividade e devidamente treinada.	8
Profissionais que realizam esta atividade e outras atividades no projeto e que trabalham isoladamente em relação aos demais administradores de dados.	6

Nenhum dos membros da empresa ou equipes.	3
Profissionais dedicados exclusivamente a esta atividade, com treinamento adequado, alocados diretamente nas equipes de projeto e que trabalham de modo colaborativo entre eles.	1
Não estou envolvido com o assunto.	1
Profissionais dedicados exclusivamente a esta atividade, com treinamento adequado, alocados diretamente nas equipes de projeto, mas que trabalham isoladamente dos demais administradores de dados.	1
12) Na sua organização foi adotado algum padrão de metadados para descrição das fontes de informação?	
Não, a empresa não segue nenhum padrão mas a descrição das fontes de informação é realizada em documentos não estruturados mantidos em um repositório sob responsabilidade da área de administração de dados.	5
Sim, a empresa segue um (ou mais de um) padrão próprio da instituição definido pelos administradores de dados.	5
não sei responder / não sei informar.	4
Não, a empresa não segue nenhum padrão e as fontes de informação não são descritas	3
Sim, a empresa segue um (ou mais de um) padrão definido por uma organização que esta faz parte (ou é legalmente vinculada) e que agrega instituições afins e internalizado pela área de administração de dados.	2
Não, a empresa não segue nenhum padrão e as fontes de informação não são descritas.	1
13) Na sua organização existe um repositório centralizado para todos os modelos de dados?	
Não, a empresa não possui um repositório centralizado e fica a critério das equipes manter os modelos atualizados.	7
Não, mas algumas equipes mantêm os modelos atualizados apesar de não ser obrigatório.	5
Sim, e as equipes mantêm este modelos sempre atualizados, pois os procedimentos são seguidos pelos administradores de dados.	4
Sim, mas algumas equipes não mantêm estes modelos atualizados, pois são obrigatórios somente na implantação do sistema.	2
Não sei informar.	2

As respostas desta seção revelaram que as necessidades de informação nas organizações que envolvem integração ainda demandam por localização e acesso isolado, requerem muitas intervenções manuais e são atendidas por soluções pontuais (*ad-hoc*).

Quanto às práticas da área de administração de dados, apesar da existência de equipes centralizadas e dedicadas a estas atividades com procedimentos definidos em 40% das organizações, estes não são seguidos por não serem obrigatórios amplamente divulgados ou adequados às metodologias de desenvolvimento de sistemas em vigor. A falta de completeza do repositório de metadados corporativo dificulta a descoberta e reuso das fontes de informação, por isso um processo de gerenciamento de metadados por todo o seu ciclo de vida é necessário para atingir a

Governança de Informação. Este processo também inclui a atualização dos modelos de dados, uma vez que estes contêm parte da semântica das fontes que nem sempre é possível expressar de modo não ambíguo através de descrições em linguagem natural.

A padronização do nome de elementos, domínios e tipos de dados se mostrou uma preocupação na maioria das organizações e mesmo naquelas onde uma norma específica não existia, a equipe de desenvolvimento do projeto criava a sua própria. As abordagens de casamento de esquemas baseadas nestas características dos elementos do esquema podem ser beneficiadas pela adoção de padrão dentro da organização, mas esforços isolados não trazem o mesmo benefício, pois ainda podem existir conflitos entre sistemas desenvolvidos por equipes diferentes.

A avaliação da abordagem e do protótipo foi positiva como pode ser observado pela tabela 5.11. Metade dos respondentes sugeriu que as recomendações de expansão sejam apresentadas através de um ou mais grafos, que é uma estrutura que permite explorar todas as ligações entre os recursos na visualização de ontologias.

Tabela 5.11 Perguntas e Respostas da Seção "*Sobre o Experimento*"

14) Os termos sugeridos pelo sistema lhe permitiram esclarecer a busca que estava sendo realizada?	
Sim, em alguns casos.	11
Sim, em todos os casos.	9
Não, em nenhum dos casos.	0
15) A apresentação dos termos e o tipo de expansão realizada pelo sistema através de uma tabela agrupada pelos termos iniciais da busca deixaram clara a relação entre os conceitos que estes termos representam?	
Sim, em todos os casos.	10
Sim, em alguns casos.	10
Não, em nenhum dos casos.	0
16) Que outra forma de apresentação seria mais adequada para representar a relação entre os conceitos e os termos sugeridos pelo sistema?	
Um ou mais Grafos	10
Uma tabela agrupada pelo tipo de expansão	8
Uma *word tree*	6
não sei informar	1
Uma *word cloud*	1

5.4 Considerações Finais sobre o Experimento

O experimento, usando um protótipo da arquitetura proposta, evidenciou que a abordagem de expansão semântica assistida pelo usuário obteve um aumento de 34,07% no que diz respeito à precisão e de 15,48% na cobertura em comparação com busca convencional por palavras-chaves. A expansão semântica automática foi descartada por não permitir que o usuário interaja com a ontologia de modo a adquirir

conhecimento sobre o domínio e melhorar a especificação de sua consulta. Enquanto que a expansão estatística automática não explora a existência das anotações semânticas no repositório, requer um esforço computacional maior para identificar os termos com maior freqüência de co-ocorrência na coleção e gerar a consulta expandida, além de recuperar um número maior de falsos positivos (baixa precisão).

A análise qualitativa do *log* de consultas permitiu identificar recursos que podem ser adicionados à ontologia de domínio através de um processo de engenharia de ontologias, de modo a torná-la mais completa e melhorar o resultado das consultas ao repositório de metadados. A evolução da ontologia é necessária para reduzir a distância entre o vocabulário de quem busca e de quem cataloga as fontes de informação. Com este processo também é possível agregar novos termos que representam neologismos, jargões e abreviações particulares do domínio como rótulos alternativos para os recursos existentes, acompanhando a evolução do vocabulário do domínio.

Esta análise também permitiu evidenciar que, apesar de os usuários terem sido orientados a informar termos que correspondem somente aos conceitos associados às necessidades de informações, alguns termos utilizados nas buscas foram mapeados em outros tipos de recursos. Esta situação sugere que a perspectiva de quem realiza a busca quanto ao que vem a ser um conceito, relacionamento, atributo ou instância nem sempre corresponde a perspectiva de quem modelou a ontologia. Neste sentido, a arquitetura compensa esta divergência ao utilizar rótulos de qualquer tipo de recurso modelado nas ontologias de domínio para realizar o casamento dos termos das consultas.

Os resultados promissores obtidos com o experimento motivaram o desenvolvimento de um estudo de caso para avaliar a aplicabilidade e melhorias que a arquitetura pode proporcionar em um caso real. Este estudo de caso será apresentado no próximo capítulo.

6. Avaliação da Proposta – Estudo de Caso

O experimento demonstrou que a busca semântica atinge melhores resultados do que a busca convencional, considerando as medidas precisão, cobertura e F1, ao explorar o conhecimento do domínio e permitir que o usuário especifique de maneira mais precisa a sua intenção de busca. Adicionalmente, a análise qualitativa do questionário revelou que a integração de informações, em algumas empresas, ainda requer que o usuário localize, realize o acesso e integre informações manualmente e que as práticas de governança da informação ainda precisam atingir níveis mais altos de maturidade para garantir a completeza do repositório de metadados.

O segundo método de avaliação (estudo de caso) visa demonstrar a viabilidade e os benefícios de aplicação da proposta em um ambiente real. Este método foi escolhido por permitir investigar um fenômeno contemporâneo dentro de seu contexto na vida real (Yin 2005) e analisar a aplicação de sistemas de informação em organizações considerando as práticas organizacionais e o conhecimento das pessoas envolvidas, para entender e explicar um fenômeno social (Myers 1997).

O Catálogo de Informações do portal *DadosGov* COI-PR [1] foi utilizado nesta fase do estudo por ser uma iniciativa do governo federal brasileiro em disponibilizar dados em formatos abertos para serem reutilizados e integrados com outras informações pelo próprio governo e pela sociedade em geral. Este catálogo foi criado a partir da coleta de séries históricas sobre indicadores dos 8 anos do Governo Lula junto aos órgãos do governo federal. Uma cópia do repositório de metadados da aplicação além de documentos, modelos, especificações e acesso à área restrita do portal foram cedidos pelo SERPRO. Reuniões presenciais e comunicações por e-mail com a equipe responsável pela aplicação foram realizadas de modo a coletar informações adicionais, dirimir as dúvidas quanto aos documentos e apresentar a arquitetura da proposta.

Este capítulo apresenta, na seção 6.1, uma breve introdução sobre Dados Abertos Governamentais para contextualização das iniciativas de criação de catálogos

[1] https://i3gov.planejamento.gov.br/

para disseminação de dados públicos. O portal *DadosGov* COI-PR é apresentado na seção seguinte (6.2), onde é descrito o processo de coleta e catalogação das séries no portal e o esquema de metadados utilizado. A seção 6.3 é dedicada a expor os aspectos da e-PING (Padrões de Interoperabilidade de Governo Eletrônico) relacionados com a interoperabilidade semântica com foco no Padrão de Metadados do Governo Eletrônico (e-PMG) e o Vocabulário Controlado do Governo Eletrônico (VCGE) que foi utilizado na criação da uma ontologia de domínio da Administração Pública. A prova de conceito construída para busca semântica por séries históricas do portal é descrita na seção 6.4, o seu potencial de uso é explorado através de um exemplo no item 6.4.4 e o resultado das buscas realizadas utilizando a ferramenta é analisado quantitativa e qualitativamente na seção 6.5. A seção 6.6 conclui o capítulo com as considerações finais sobre o estudo de caso.

6.1 Dados Abertos Governamentais

Organizações públicas são os maiores produtores e consumidores de informações. As informações públicas são produzidas, arquivadas e divulgadas de várias maneiras e formatos, desde veículos de publicação em papel como em Diários Oficiais até sítios na internet, no formato que for mais conveniente para o órgão governamental que gerou a informação e de acordo com as normas vigentes (W3C Brasil 2011b).

Iniciativas recentes, de países como os Estados Unidos e Inglaterra, foram criadas para promover a publicação sistemática de informações de ações governamentais. De acordo com a W3C (W3C Brasil 2011a) estas iniciativas, conhecidas como Dados Abertos Governamentais, têm por objetivo "*a publicação e disseminação das informações do setor público na Web, compartilhadas em formato bruto e aberto, e compreensíveis logicamente, de modo a permitir a sua reutilização em aplicações digitais desenvolvidas pela sociedade*".

A publicação de dados governamentais abertos requer aderência aos princípios de dados abertos conforme definido por (Malamud *et. al.* 2007): (1) **Completos**: todos os dados públicos devem ser disponibilizados, (2) **Primários**: os dados devem ser apresentados tais como coletados na fonte (formato bruto), ou seja, com o maior nível possível de detalhamento e sem agregação ou modificação, (3) **Atuais**: o tempo de publicação deve ser adequado à preservação do seu valor, (4) **Acessíveis**: a disponibilização dos dados deve atingir o maior número possível de usuários, (5) **Compreensíveis por máquinas**: através de uma estrutura que permita o seu processamento, (6) **Não discriminatórios**: não requerem cadastro ou formalização de

pedido para acesso, (7) **Não proprietários**: o formato aberto é usado para armazenar e disseminar dados digitais, pois é livre de limitações legais quanto ao seu uso e (8) **Livre de licenças**: os dados não têm restrições de uso quanto a direitos autorais, patentes ou outras restrições legais e também não violam a privacidade individual ou comprometem a segurança.

Catálogos de dados, como o DATA.GOV (Estados Unidos), data.gov.uk (Reino Unido) e data.worldbank.org (do Banco Mundial), foram criados nos anos de 2009 e 2010 com o objetivo de tornar mais fácil a localização e uso pelo o público em geral dos dados governamentais publicados. Os dados e seus metadados disponibilizados através destes catálogos permitem que outros agentes (públicos ou privados) misturem, melhorem e compartilhem essas informações, aumentando a integração de dados e proporcionando o surgimento de novos serviços (W3C Brasil 2011b).

Além da publicação também é necessário avaliar como os dados estão sendo efetivamente reutilizados. Alguns concursos foram realizados para descobrir quais são os aplicativos mais procurados, como *Show Us a Better Way* no Reino Unido e o *Apps for Democracy* dos Estados Unidos enquanto outros governos realizam consultas públicas sobre acesso aberto a informações públicas como a do governo da Austrália (W3C Brasil 2011b).

A reutilização de dados governamentais, inclusive através da integração com dados de outras fontes de dados, requer que a semântica destas informações seja estabelecida, de modo preciso e explícito, e associada aos dados publicados (Harris *et. al.* 2008). Caso contrário, dados referentes a conceitos que não são semanticamente equivalentes e nem mesmo relacionados podem ser integrados, gerando resultados errôneos.

A fim de evitar este tipo de problema, além dos dados em si, devem ser publicados os seus metadados e os conceitos e definições do vocabulário específico do domínio, em formato partilhável e referenciável, de modo a contextualizar e transformar os dados em informações. Através das tecnologias desenvolvidas pelo avanço da Web Semântica, os dados governamentais podem ser disponibilizados anotados em relação a este vocabulário e as interfaces de busca podem permitir que os aplicativos recuperem e acessem estas informações de uma forma não pré-definida (W3C Brasil 2011b).

Neste sentido algumas iniciativas do governo brasileiro já realizam a disseminação de dados públicos em formato aberto associados a seus metadados, como é o caso do portal *DadosGov* COI-PR que será apresentado na próxima seção. De modo a garantir que estas iniciativas independentes estejam em conformidade com a e-PING e atendam às condições de disseminação e compartilhamento de dados e

informações públicas no modelo de Dados Abertos, foi criada a Infraestrutura Nacional de Dados Abertos (INDA), que teve como inspiração a Infraestrutura Nacional de Dados Espaciais (INDE) [2]. A INDA tem por objetivo: "*(1) proporcionar a busca, o acesso, o reuso e o cruzamento dos dados públicos de diferentes fontes e assuntos de maneira simples e eficiente, (2) coordenar e orientar a padronização na geração, armazenamento, acesso, compartilhamento, e disseminação dos dados e informações públicas de governo e (3) incentivar a agregação de valor aos dados públicos e fomentar a colaboração com o cidadão na implementação de novos serviços à sociedade*" (SLTI-MP 2011).

6.2 *DadosGov* COI-PR

O portal *DadosGov* COI-PR contém um catálogo de mais de 1900 séries históricas sobre indicadores de resultados de ações, projetos e programas do governo federal. Destas, mais de 1100 estão disponíveis para acesso ao público em geral, mas aproximadamente 800 ainda estão em processo de coleta, carga ou catalogação dos dados. Por exemplo, o catálogo contém séries como indicadores do número de pessoas beneficiadas e do valor anual investido em programas de transferência de renda distribuídos por estado, como o programa Bolsa Família.

Através da análise de planilhas, documentos, apresentações e pequenos bancos de dados utilizados pelo Comitê de Organização de Informações da Presidência da República (COI-PR) e fornecidos por órgãos que compõem a estrutura do governo federal foi possível identificar que (1) a maior parte das informações utilizadas no apoio à gestão pública de nível estratégico é apresentada em três dimensões: temporal, espacial e o órgão responsável pelos dados e (2) os indicadores (também chamados de variáveis) podem ser agrupados em função de assuntos em comum.

O processo de coleta e catalogação dos dados das séries históricas envolve o COI-PR (denominado Cliente Central), os órgãos responsáveis pelos dados (denominados Clientes Setoriais) e o SERPRO conforme as atividades apresentadas na figura 6.1. O estágio da série histórica é atualizado até que o processo seja completado e a série se torne disponível para o público. Os metadados do Grupo de Informação são obrigatórios a partir do estágio "Dados Solicitados" enquanto que os metadados das séries a partir do estágio "Dados no Formato Original", mas ambos os

[2] http://www.inde.gov.br/

conjuntos de metadados podem ser alterados nos estágios subseqüentes pelo grupo de usuários do COI-PR.

O esquema de metadados definido para catalogação das séries históricas, apresentado nas tabelas 6.1 e 6.2, e o modelo físico do banco de dados centralizado para armazenamento dos dados e metadados (figura 6.2) dos grupos de informação (temática) e das séries (variáveis) refletem estas características dos dados e do processo.

Tabela 6.1 Metadados das Séries Históricas

Elemento	Semântica do elemento
Estágio de preparo das séries históricas	Estágio de preparo por que passam as séries históricas desde o formato original, recebido dos Órgãos Gestores dos Dados, até a sua publicação na Web. Os estágios são os seguintes: original, padrão, armazenado, produção, validado e publicado.
Grupo Informação	Nome do Grupo de Informação. Conceito que organiza as séries históricas por um domínio de negócio. É utilizado principalmente para organizar as séries históricas a serem enviadas pelos órgãos Gestores.
ID Série Histórica	Código da série histórica no banco de dados.
Nome Série Histórica	Nome da série histórica.
Nome Reduzido	Por restrições da ferramenta de tratamento dos dados o nome da série histórica precisar ser reduzido para o limite de 130 caracteres e deve ser único entre todas as séries.
Descrição Série Histórica	Descrição detalhada da série histórica.
Produto	As séries históricas apresentam dados financeiros e quantitativos das ações de Governo. Produto é o resultado efetivo da Ação de Governamental valorada na série histórica. Exemplos de alguns valores permitidos: Pessoas beneficiadas, Pessoas capacitadas, Valor captado, Valor contratado, Valor investido, Valor repassado.
Unidade Medida	Em qual unidade se expressa os valores apresentados nas séries históricas. Exemplos de alguns valores permitidos: Pessoas, Famílias, Hectares, R$, Tonelada Equivalente de Petróleo (TEP).
Multiplicador	Se o valor apresentado deve ser lido multiplicado. Valores permitidos: em mil, em milhão, em bilhão, nulo.
Fonte Gestora	Órgão ministerial responsável pela gestão dos valores informados na série histórica.
Fonte Provedora	Órgão ministerial responsável pelo fornecimento da série histórica.
Órgão Primeiro Escalão	Classificação das Fontes Gestoras conforme componente do primeiro escalão de Governo.
Início Série	Ano ou ano, mês do início da série histórica.
Final Série	Ano ou ano, mês do final da série histórica.
Armazenamento	Formato do valor apresentado na série histórica. Valores permitidos: número inteiro, real com duas casas decimais, real com quatro decimais.
Aditividade Local	Se os valores das séries históricas referentes a município podem ser consolidados para UF e Brasil e se de UF podem ser

	consolidados para Brasil.
Aditividade Tempo	Se os valores das séries históricas referentes a mês podem ser consolidados para ano.
Página de origem dos dados	Sítio de onde foram extraídos os dados, caso estejam na Web.
Data Atualização	Data da última atualização da série histórica no banco de dados.
Temática	Classificação temática a partir de taxonomias de assuntos próprias da aplicação. Um exemplo é a taxonomia “Balanço de 8 Anos”. Este elemento tem uma ocorrência obrigatória e permite várias ocorrências.

Tabela 6.2 Metadados dos Grupos de Informação

Elemento	Semântica do elemento
ID Grupo Informação	Código do Grupo de Informação no banco de dados.
Grupo Informação	Nome do Grupo de Informação. Conceito que organiza as séries históricas por um domínio de negócio. É utilizado principalmente para organizar as séries históricas a serem enviadas pelos órgãos Gestores.
Tipo do Grupo Informação	Se as séries históricas do grupo tem acesso público (sociedade) ou restrito (somente pessoas autorizadas). Valores permitidos: público, restrito.
Periodicidade	As séries históricas apresentam valores por ano ou mês. Valores permitidos: anual, mensal.
Base Territorial	As séries históricas apresentam dados consolidados por municípios, estados ou nacionais. Valores permitidos: municipal, estadual, nacional.

O comitê está realizando também um levantamento junto aos clientes setoriais, através de um questionário, de modo a coletar informações adicionais sobre as séries históricas. O levantamento visa identificar a aderência aos princípios de dados abertos e a qualidade dos dados contidos nas mesmas. Através deste levantamento será possível adicionar outros metadados às séries publicadas como os dados de contato do responsável pelas mesmas na fonte gestora, o sistema de origem dos dados da série, se os dados são arquivados neste sistema, se são publicados em algum outro sítio na internet e se passam por processos de auditoria.

Em caso de séries secundárias, ou seja, derivadas de outras séries, este levantamento visa identificar as séries históricas primárias, os cálculos e regras de transformação aplicadas nas séries primárias para gerar a série secundária. A cobertura espacial e temporal das séries poderá ser medida a partir do registro da legislação pertinente às ações de governo relacionadas com os indicadores da série, de eventos que impliquem em descontinuidade ou mudança metodológica na coleta das informações e de justificativa para ausência de valores nas dimensões de espaço e tempo.

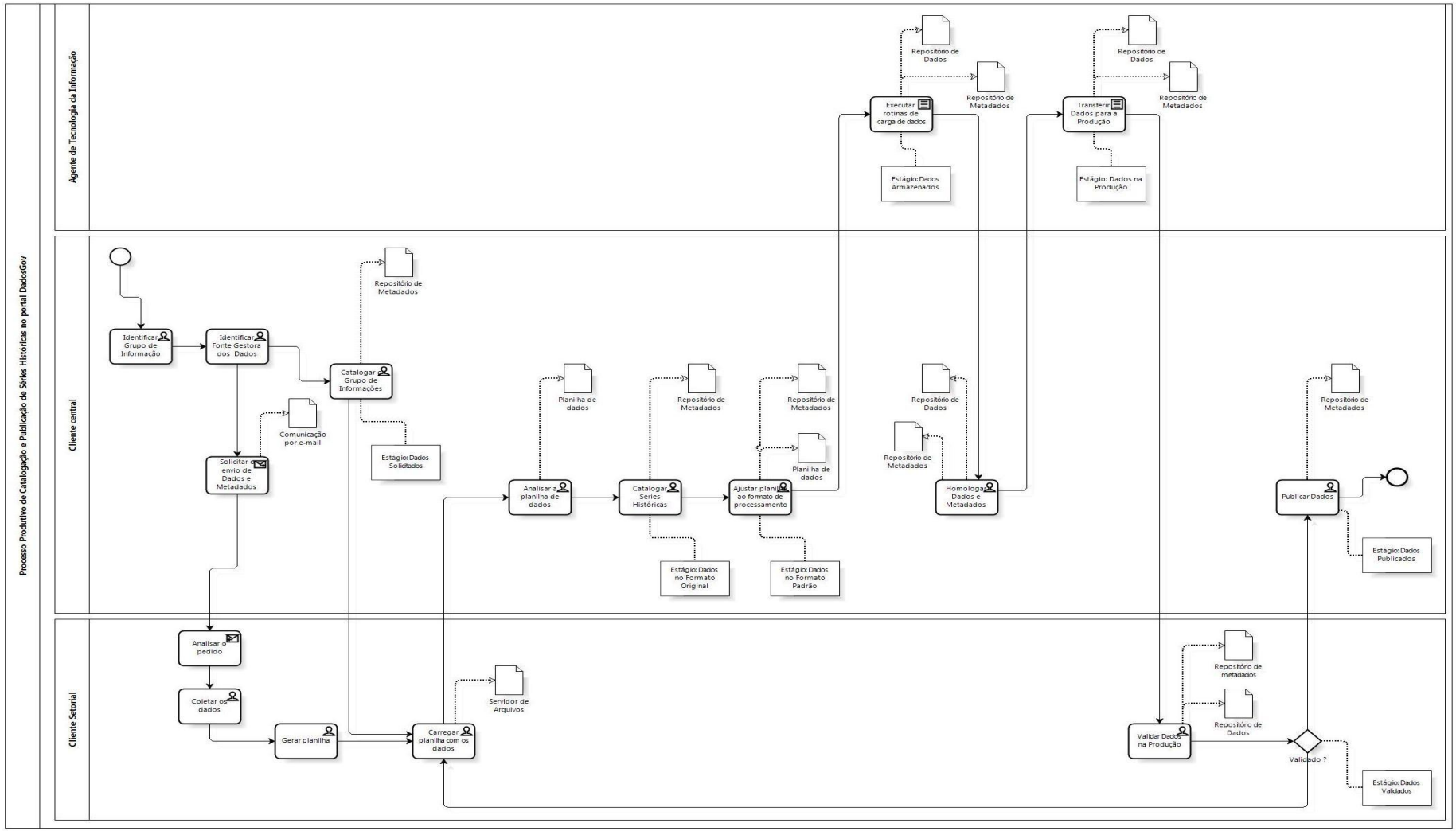


Figura 6.1 Processo Produtivo de Coleta e Catalogação de Dados no Portal

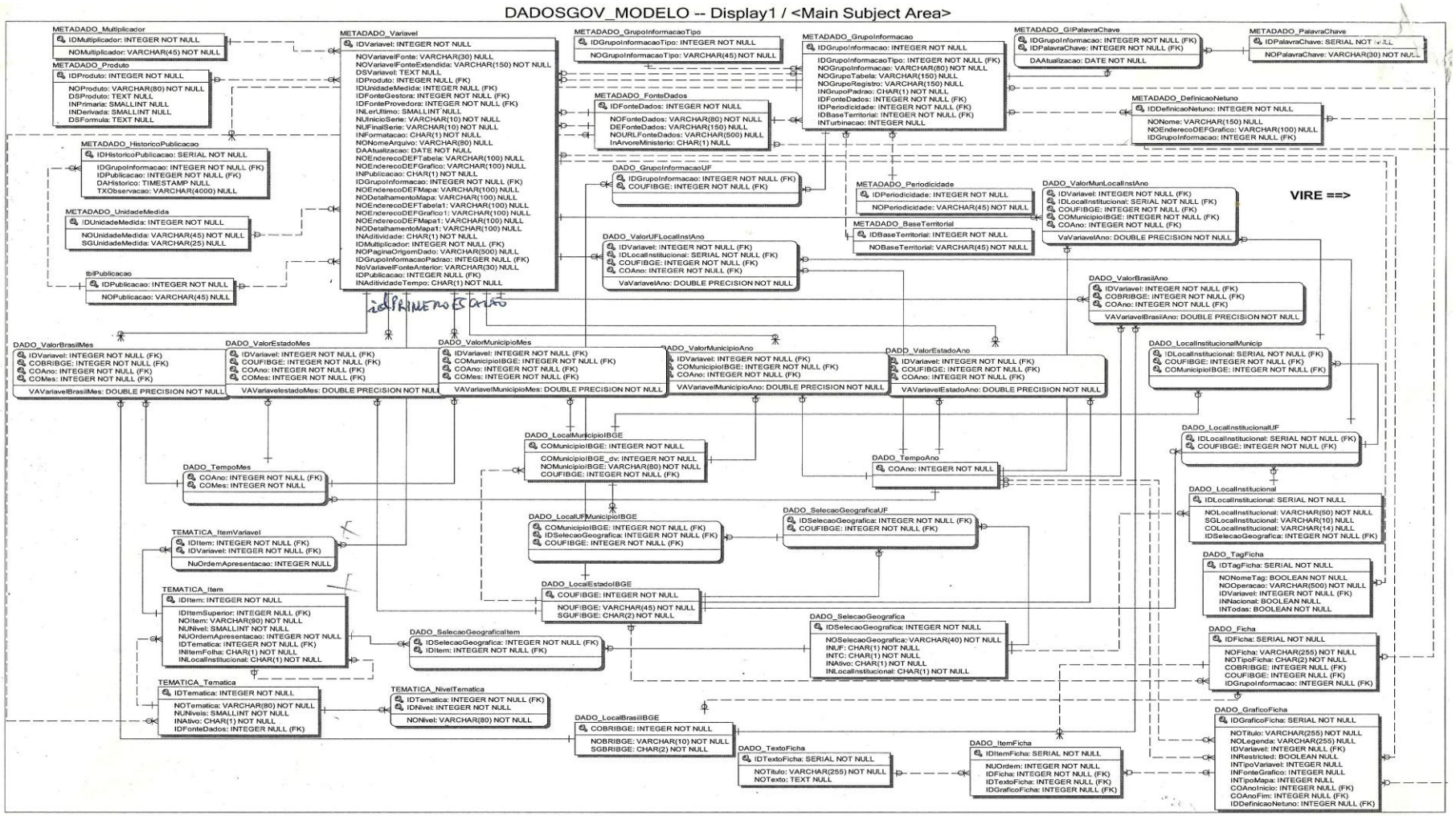


Figura 6.2 Modelo Físico do Repositório de Dados e Metadados

A solução de integração de informações desenvolvida para o portal *DadosGov* COI-PR segue a abordagem de armazenamento comum de dados (Ziegler e Ditrich 2007) pois estes são extraídos de suas fontes originais e armazenados em um repositório central único no formato bruto de acordo com nível de granularidade temporal e espacial do seu fornecimento. Apesar do modelo de dados não seguir a modelagem dimensional clássica, não utilizando tabelas fatos e dimensões, esta solução possui muitas características em comum com um DW por realizar a integração materializada de informações agrupadas por assuntos, através do elemento Grupo Informação, não voláteis e associadas à dimensão temporal para apoio a decisão do nível estratégico, neste caso a Presidência da República. A analogia com ambiente de DW também se aplica às possibilidades de consultas ao portal, típicas de processamento analítico (OLAP – On Line Analytical Processing).

Através do portal *DadosGov* COI-PR, qualquer cidadão pode realizar as seguintes operações com as séries que se encontram no estágio publicado: (1) visualizar os dados através de tabelas, gráficos e mapas, associados aos metadados, de acordo com filtros espaciais e temporais, (2) realizar o *download* dos dados em formatos abertos [3] (CSV, RTF, PDF e XML), (3) compartilhar o resultado de consultas por e-mail, (4) recuperar notícias, relacionadas com as séries históricas, publicadas no portal por agências de notícias oficiais e (5) buscar séries históricas de acordo com as árvores temáticas associadas ao Grupo de Informação.

Considerando o investimento empregado, do ponto de vista do tempo e esforço, para a publicação destes indicadores no portal, é importante avaliar outras formas de recuperação dos dados, principalmente por parte da sociedade em geral que não está familiarizada com as árvores temáticas que foram usadas para classificação das séries históricas.

Uma prova de conceito da arquitetura proposta por esta pesquisa foi disponibilizada na internet para permitir a busca por palavras-chave no conteúdo dos metadados do Catálogo de Informações e a especificação de critérios de qualidade a serem aplicados na ordenação dos resultados. A aplicação construída, que será detalhada na seção 6.4, utilizou uma ontologia de domínio gerada a partir do Vocabulário Controlado de Governo Eletrônico (VCGE) proposto pela arquitetura e-PING, que será apresentado na próxima seção.

[3] Formato aberto é todo formato cuja especificação está disponível publicamente sem restrições para uso para permitir a interoperabilidade de softwares. Fonte: http://www.openformats.org/enShowAll

6.3 e-PING

A arquitetura e-PING estabelece um conjunto mínimo de premissas, políticas e especificações técnicas com o propósito de nortear as condições de interação entre os órgãos do governo federal e com a sociedade em geral. Estes padrões visam assegurar que sistemas, processos e informações sejam gerenciados de modo a maximizar oportunidades de troca e reuso de informações (e-PING 2010). O documento está segmentado em cinco grupos e um destes grupos, denominado de Organização e Intercâmbio de informações, é responsável por avaliar e definir os padrões para disponibilização e troca de informações públicas. Dentre estes padrões estão o Padrão de Metadados do Governo Eletrônico (e-PMG) e o Vocabulário Controlado do Governo Eletrônico (VCGE).

6.3.1 e-PMG

O e-PMG (e-PMG 2010) é um esquema de metadados que estabelece um conjunto mínimo de elementos necessários para a recuperação e gerenciamento de informações públicas, que podem estar em meios eletrônicos ou não. Este esquema foi desenvolvido com base no Dublin Core (DC) e em outros esquemas de metadados como o e-GMS (*Government Metadata Standard*) do governo do Reino Unido. O e-PMG consiste em 20 elementos: os 15 elementos do DC (Abrangência, Assunto, Colaborador, Criador, Data, Descrição, Direitos, Fonte, Formato, Identificador, Idioma, Publicador, Relação, Tipo e Título) e 5 elementos adicionais (Contexto jurídico-administrativo, Destinação, Destinatário, Localização e Preservação) identificados como necessários para o contexto do governo eletrônico brasileiro.

A expectativa do governo federal com a adoção do e-PMG é que a sua utilização permita que os cidadãos localizem de modo eficiente as informações governamentais que desejam sem necessidade de um conhecimento detalhado da estrutura governamental. Espera-se também o aumento do compartilhamento de informações e serviços entre os órgãos governamentais (government-to-government ou G2G) e entre governo e sociedade (government-to-citizen ou G2C), o que resultaria em maior eficiência na gestão de informações, na transparência do governo e na participação dos cidadãos.

O documento de referência do e-PMG indica os elementos obrigatórios, possíveis qualificadores e esquemas de codificação para alguns elementos. Quanto ao momento de criação dos metadados, o padrão recomenda que estes sejam criados o mais cedo possível no ciclo de vida do recurso informacional, por quem for responsável pela criação ou publicação do mesmo. Mas, de acordo com a última

versão do documento de referência da e-PING (e-PING 2010) este padrão é considerado como “Em Estudo”, ou seja, a sua utilização ainda não é obrigatória.

O elemento Assunto deste esquema de metadados tem por finalidade a especificação de palavras-chave ou termos que representem corretamente o conteúdo do recurso para possibilitar a recuperação através da busca por palavras-chave ou navegação em diretórios de classificação. Este elemento admite três qualificadores: Categoria, Palavra-chave e Pessoa. O qualificador Categoria pretende apoiar a navegação por um diretório de classes mais amplas, enquanto que o qualificador Palavra-chave apóia a pesquisa direta, mas os valores para todos os qualificadores do elemento Assunto devem ser retirados de esquemas codificados como vocabulários controlados. Para o qualificador Categoria (*assunto.categoria*) a recomendação é a utilização do VCGE enquanto que para o qualificador Palavra-chave (*assunto.palavra-chave*) podem ser utilizados tesauros específicos que sejam pertinentes a natureza do recurso ou um vocabulário controlado consagrado no mercado. O terceiro qualificador é aplicado para especificar quando o recurso informacional é sobre uma pessoa (*assunto.pessoa*).

6.3.2 VCGE

O VCGE (VCGE 2010) é um vocabulário controlado sobre assuntos da Administração Pública, que segue a norma ISO 2788 para criação de tesauros em um único idioma. A especificação deste esquema de codificação tem como foco o cidadão, por isso este contém termos do vocabulário do cidadão como preferenciais e é independente da estrutura governamental para que não sofra impactos de mudanças de estruturas e organogramas. Este vocabulário foi criado a partir da Lista de Assuntos de Governo (LAG) que exercia este mesmo papel de esquema de codificação do elemento assunto na versão anterior da e-PING.

Este vocabulário contém termos para identificar conceitos ou um conjunto de objetos que são o referencial para a classificação de recursos informacionais gerados ou consumidos pelo governo. A primeira versão (vigente em 2011) possui 17 conceitos de primeiro nível (*top term*) e 1505 conceitos estruturados como uma taxonomia (*narrower term / broader term*) e associados aos termos indicados para uso (*preferred terms*) no preenchimento do elemento *assunto.categoria.* Vinte e cinco destes conceitos também possuem termos sinônimos que estão em desuso (*non-preferred terms*). Esta versão do vocabulário não possui outras relações entre os conceitos (*related term*) e nem descrições (*scope note*) para melhor elucidar o significado dos termos.

De acordo com a última versão do documento de referência da e-PING (e-PING 2010) este padrão é considerado como “Adotado”, ou seja, a sua utilização é obrigatória para descrição de novos recursos (a serem publicados) e também fortemente recomendada para a descrição de recursos existentes. Além de facilitar a catalogação e busca por recursos informacionais, os vocabulários controlados também podem ser usados para anotação de dados e assim permitir a integração de dados de sistemas diferentes (W3C Brasil 2011b).

6.4 Prova de Conceito

A W3C, através de seu grupo de interessem em Governo Eletrônico, estimula a interação entre os governos e os pesquisadores para a aplicação de tecnologias mais avançadas para integração de dados. Esta interação pode ocorrer através do desenvolvimento de aplicações para provas de conceito que usem informações governamentais reais, que provem as vantagens da integração de dados e que demonstrem avanços nas áreas em que governantes e cidadãos realmente têm necessidade (W3C Brasil 2011b). Neste sentido, uma prova de conceito com as funcionalidades presentes na arquitetura proposta foi construída usando o repositório de metadados do portal *DadosGov* COI-PR e para o desenvolvimento desta aplicação foram necessárias:

1) A análise do esquema de metadados de catalogação das séries no portal para identificar os elementos descritivos do seu conteúdo e os elementos referentes a características de qualidade das séries históricas.
2) Uma ontologia de domínio que modelasse o contexto da Administração Pública Federal. A ontologia utilizada foi modelada a partir do VCGE.
3) A Anotação semântica de um subconjunto das séries históricas usando os conceitos do VCGE.
4) A adaptação do protótipo usado no experimento com as funcionalidades de busca e registro de consultas para recuperar os metadados das séries históricas e sua disponibilização na internet.

6.4.1 Análise do Esquema de Metadados do Portal

A análise do esquema de metadados permitiu identificar cinco elementos para aplicação da busca textual com a lista de palavras-chave semanticamente expandida, que são: Nome completo, Descrição, Grupo Informação, Temática e Produto. Os elementos Nome Completo e Descrição correspondem a texto livre, mas os elementos

Grupo Informação, Temática e Produto possuem uma lista de valores predefinida para cada um. O elemento Nome Reduzido não foi utilizado por não ter por objetivo a descrição do recurso para fins de localização e sim uma identificação a ser usada pela ferramenta por restrições técnicas.

A função do elemento Grupo Informação é agrupar séries afins que fornecem indicadores diferentes referentes ao mesmo assunto (*o quê?*). O repositório de metadados utilizado na prova de conceito possui 415 grupos cadastrados e, no portal *DadosGov* COI-PR, através do Índice do Catálogo é possível ter acesso às séries de cada grupo.

Alguns exemplos são apresentados na tabela 6.3 e a análise destes exemplos permite identificar que o grupo informação também agrupa séries de acordo: (1) com a fonte gestora (Resíduos Sólidos: *quem?* FUNASA x Ministério das Cidades), (2) com a fonte provedora dos dados (Indicadores sociais: *quem?* Ipeadata), (3) com o nível de granularidade das séries (Economia solidária: *onde?* nos estados), (4) com o programa de governo (Minha Casa Minha Vida, Reuni) ou suas modalidades (Projovem), (5) com assuntos mais específicos (Formação de professores X Formação de professores na educação especial) e (6) com assuntos similares e disjuntos (Saneamento em áreas especiais X Saneamento em áreas indígenas).

Alguns assuntos podem ser mapeados diretamente em conceitos do VCGE (software livre, SAMU, transferência de renda) e outros em conceitos mais específicos que poderiam ser acrescentados ao vocabulário (Rodovias Federais como termo mais específico para Infraestrutura de transporte rodoviário e Indicadores Sociais como termo mais específico para Informações estatísticas).

Tabela 6.3 Exemplos de Grupos de Informação das Séries Históricas

Grupo Informação	Base Territorial	Periodicidade
Abastecimento de água e esgotamento pelo MCidades	Estadual	Anual
Abastecimento de água' pela Funasa	Estadual	Anual
Economia solidária	Nacional	Anual
Economia solidária nos estados	Estadual	Anual
Formação de professores	Estadual	Anual
Formação de professores na educação especial nos estados	Estadual	Anual
Indicadores sociais (Ipeadata)	Estadual	Anual
Minha Casa Minha Vida	Estadual	Anual
Projovem Adolescente	Municipal	Anual
Projovem Agente Jovem	Municipal	Anual
Projovem Campo	Estadual	Anual
Projovem Prisional	Estadual	Anual
Projovem Trabalhador - elevação da escolaridade e qualificação profissional	Estadual	Anual
Projovem Urbano	Municipal	Anual
Resíduos Sólidos – Funasa	Estadual	Anual
Resíduos Sólidos – Mcidades	Estadual	Anual

Reuni - Reestruturação e Expansão das Universidades Federais	Estadual	Anual
Rodovias Federais	Estadual	Anual
Samu - Serviço de Atendimento Móvel de Urgência	Estadual	Anual
Saneamento em áreas especiais pela Funasa	Estadual	Anual
Saneamento em áreas indígenas pela Funasa	Estadual	Anual
Software livre	Estadual	Anual
Transferência de renda	Municipal	Anual

Um Grupo de Informação também pode corresponder a um conceito de uma ou mais taxonomias temáticas. As seguintes taxonomias estão cadastradas no repositório de metadados do portal: COI-PR, LAG (e-PING), Previdência, Holograma, PPA, Funcional Programática, Balanço de 8 anos e Ministérios. A tabela 6.4 apresenta a quantidade de séries associadas a cada taxonomia. Atualmente a aplicação permite a navegação somente pelas taxonomias COI-PR e Balanço de 8 anos.

A utilização deste elemento do esquema de metadados na busca textual considerou todas as suas ocorrências e todos os níveis superiores do grupo aos quais a série estava associada. Por exemplo, na série "Valor conveniado pelo MEC para construção/reforma/ampliação de escolas indígenas" (id 2009), o elemento possui três ocorrências: (1) Índice do Catálogo > Escolas indígenas, (2) Balanço de 8 anos > Cidadania e Inclusão Social > Educação > Escolas indígenas e (3) Ministérios > Ministério da Educação > Escolas indígenas.

Tabela 6.4 Taxonomias de classificação das Séries Históricas

Taxonomia	Número de séries classificadas
Previdência	3
LAG (e-PING)	10
Holograma	35
PPA	50
Funcional Programática	55
COI-PR	231
Balanço de 8 anos	1869
Ministérios	1878

O elemento Produto descreve o tipo de indicador contido na série. Este elemento permite especificar, com diferentes níveis de precisão, o que o indicador representa, cabendo ao responsável pela catalogação dos metadados a decisão pelo nível de precisão utilizado. A lista de valores possíveis restringe o conteúdo deste elemento a um esquema de codificação e evita problemas de ambigüidade terminológica. Mas não estabelece uma diferenciação entre as categorias existentes para apoiar o responsável pela catalogação nesta escolha. A descrição dos conceitos transformaria este esquema em um glossário e permitiria, por exemplo, identificar

quando utilizar “Alunos matriculados” e “Matrículas realizadas” ou "Família assistida", "Família beneficiada" e "Famílias Atendidas".

Quatro elementos do esquema de metadados contêm informações que representam características de qualidade das fontes de informação no contexto de dados abertos governamentais: Fonte Gestora, Fonte Provedora, Base Territorial e Periodicidade. A Fonte Gestora e a Fonte Provedora especificam a origem dos dados da série. A Base Territorial especifica o nível de granularidade espacial dos dados e enquanto que a Periodicidade especifica o nível de granularidade temporal. No portal existem 6 séries Mensais e 1918 Anuais, 136 com dados por Município, 1284 por Estado e 504 Nacionais e 15 séries nas quais a fonte provedora dos dados da série é diferente da fonte gestora.

Conforme os oito princípios para divulgação de Dados Abertos Governamentais, estes devem ser fornecidos no maior nível possível de detalhamento por isso a granularidade temporal e espacial foram transformadas em medidas ordinais para indicar que a granularidade da Periodicidade Mensal é menor que a Anual e que a da Base Territorial Municipal é menor que a Estadual e esta é menor que a Nacional. Proveniência de dados descreve a origem dos dados e o processo que transfere estes dados da fonte original para o repositório onde está armazenado. O processo de publicação das séries no portal, apresentado na figura 6.1, é padronizado para qualquer fonte, então a qualidade da proveniência dos dados foi considerada melhor quando a Fonte Gestora for igual à Fonte Provedora.

Para o estudo de caso, um novo elemento, chamado Anotação Semântica, foi adicionado ao esquema de metadados das séries históricas. Este elemento possui a mesma semântica do qualificador *Requires* aplicada ao elemento *Relation* do padrão Dublin Core (DCMI 2010). O conteúdo recomendado pelo DC para este elemento é uma string ou número em conformidade com um sistema formal de identificação como, por exemplo, o URI. O objetivo é referenciar um recurso da ontologia de domínio através do seu URIref a partir da série histórica que foi anotada. O qualificador *Requires* identifica uma relação, onde o recurso descrito (neste caso, a série histórica) requer o recurso referenciado (neste caso um conceito ou instância que pertence à ontologia de domínio) para suportar a sua função, entrega e coerência do conteúdo. A inclusão deste elemento acrescenta semântica formal ao esquema de metadados do catálogo, aumentando o seu potencial de explicação e localização das fontes de informação, assim como permite a identificação das fontes que possuem conceitos semanticamente similares ou pelo menos relacionados usando uma ontologia de domínio como referência.

Para o estudo de caso, uma ontologia formal leve, codificada na linguagem OWL-DL, foi modelada a partir do VCGE (ontologia informal). Cada termo do tesauro foi convertido em um conceito com um rótulo em linguagem natural. As relações existentes entre os termos como mais geral e mais específico (TG/TE) refletiu na estrutura hierárquica da ontologia com conceitos pais e conceitos filhos. Os termos em desuso (UP) também foram adicionados como rótulos dos conceitos uma vez que, diferente dos vocabulários controlados, a ontologia de domínio não exerce a função de controle terminológico.

6.4.2 Anotação de Séries Históricas usando o VCGE

O casamento entre os rótulos da ontologia e o conteúdo dos elementos descritivos selecionados do esquema de metadados foi utilizado como fonte para sugestão de anotações para as séries e algumas destas sugestões foram analisadas individualmente. A tabela 6.5 apresenta alguns exemplos deste casamento usando a função *ts_tank* para gerar a pontuação de semelhança e a indicação se a sugestão foi usada ou não para anotação.

Tabela 6.5 Anotações sugeridas através do casamento dos rótulos da ontologia com os metadados descritivos selecionados

Conceito	Rótulo	ID da série	Pontuação	Anotada
#Software_livre	Software livre	1746	0.874054	Sim
#Software_livre	Software livre	1749	0.874054	Sim
#Software_livre	Software livre	1747	0.874054	Sim
#Software_livre	Software livre	1745	0.874054	Sim
#Rodovia_Federal	rodovia federal	158	0.862973	Sim
#Rodovia_Federal	rodovia federal	157	0.862973	Sim
#Rodovia_Federal	rodovia federal	160	0.861023	Sim
#Rodovia_Federal	rodovia federal	162	0.853803	Sim
#Rodovia_Federal	rodovia federal	156	0.849861	Sim
#Rodovia_Federal	rodovia federal	161	0.847662	Sim
#Rodovia_Federal	rodovia federal	159	0.841125	Sim
#Petroleo	Petróleo	212	0.0919062	Sim
#Petroleo	Petróleo	253	0.0759909	Sim
#Petroleo	Petróleo	358	0.0759909	Sim
#Petroleo	Petróleo	1848	0.0759909	Sim
#Gas_natural	Gás natural	212	0.750161	Sim
#Gas_natural	Gás natural	1849	0.270901	Sim
#Gas_natural	Gás natural	253	0.198393	Sim
#Saude_bucal	saúde bucal	20	0.986179	Sim
#Saude_bucal	saúde bucal	21	0.981785	Sim
#Saude_bucal	saúde bucal	19	0.978433	Sim
#Saude_bucal	saúde bucal	326	0.749126	Sim
#Livro_didatico	Livro didático	186	0.932204	Sim
#Livro_didatico	Livro didático	181	0.887352	Sim
#Livro_didatico	Livro didático	185	0.866836	Sim
#Livro_didatico	Livro didático	184	0.866836	Sim
#Livro_didatico	Livro didático	187	0.840175	Sim
#Livro_didatico	Livro didático	180	0.833488	Sim
#Livro_didatico	Livro didático	182	0.833488	Sim

#Livro_didatico	Livro didático	179	0.833488	Sim
#Livro_didatico	Livro didático	183	0.828875	Sim
#Livro_didatico	Livro didático	757	0.158966	Não
#Livro_didatico	Livro didático	758	0.158966	Não
#Livro_didatico	Livro didático	759	0.00042839	Não

Este casamento também permitiu identificar que não foram encontradas séries históricas que pudessem ser mapeadas diretamente a 1067 conceitos da ontologia, porém isto não confirma a inexistência de fontes de informação para estes conceitos no repositório de metadados, uma vez que o VCGE não foi utilizado como um vocabulário de referência para catalogação das séries.

As dez séries que haviam sido classificadas como "*Transferência de Renda*" na Temática LAG (tabela 6.4) foram analisadas individualmente. Estas séries foram escolhidas, pois a LAG é a origem do VCGE e esta associação permitiria a anotação destas séries com o conceito **#Programas_de_Transferência_de_Renda**. Os indicadores destas séries forneciam dados sobre "Valor Repassado" e "Famílias Atendidas" (Produto) referentes a cinco programas de transferência de renda, que são: Bolsa Família, Cartão Alimentação, Bolsa Escola, Auxílio Gás e Bolsa Alimentação. Ao invés da anotação direta com o conceito, o que dificultaria a diferenciação entre os programas caso o usuário desejasse recuperar indicadores de somente um deles, foram criadas instâncias para cada um dos cinco programas na ontologia de domínio e a anotação nestas dez séries foi realizada em relação a estas instâncias.

Várias outras séries do portal também contêm indicadores de resultados de ações governamentais realizadas através de programas de governo. Considerando esta característica, outros quatro programas, identificados a partir do elemento Grupo Informação, foram criados como instâncias na ontologia e as séries destes foram anotadas com os ponteiros (URIrefs) destas instâncias. São eles: (1) "***Minha Casa Minha Vida***", instância de **#Financiamento_Habitacional** e 26 séries anotadas, (2) "***BPC - Benefício da Prestação Continuada***", instância de **#Beneficio_assistencial_ao_idoso_e_ao_deficiente** (conceito filho de **#Previdencia_social**) e quatro séries anotadas, (3) "***Luz para Todos***", instância de **#Energia_eletrica** (filho de **#Recursos_energeticos**) e oito séries anotadas e (4) "***Pronasci - Programa Nacional de Segurança Pública com Cidadania***", instância de **#Seguranca_Publica** e 17 séries anotadas.

A versão final da ontologia usada para a prova de conceito continha 1479 conceitos (classes), 9 instâncias (indivíduos), 2 relacionamentos (propriedades de objetos) e 3 atributos (propriedade de dados). Além disto, 250 das 1924 séries históricas catalogadas no repositório de metadados foram anotadas manualmente.

6.4.3 Detalhamento da Aplicação

O banco de dados *PostgreSQL* 8.4, a exemplo do experimento, é o SGBD utilizado pelo portal para o repositório de metadados e dados das séries históricas, por isso as alterações realizadas no protótipo para construção da aplicação da prova de conceito de busca semântica envolveram principalmente a adaptação ao esquema de metadados. A ontologia de domínio também foi convertida no formato RDF / OWL Database e armazenada no banco de dados. A funcionalidade *Full Text Search* foi utilizada para busca textual na ontologia e no conteúdo dos elementos selecionados do esquema de metadados.

A tela inicial da busca por séries históricas apresentada na figura 6.3 permite que o usuário descreva a sua necessidade de informação, informe até seis termos para a busca, a opção de combinação destes termos (TODOS ou QUALQUER UM) para geração da consulta executada no repositório de metadados e a especificação da prioridade em relação aos critérios de qualidade para ordenação dos resultados. O usuário também pode alterar a distância semântica máxima para busca com propagação na ontologia (0, 1 ou 2) e a quantidade máxima de séries a serem recuperadas com a busca (5, 10, 20, 50, 100, 200).

Nesta prova de conceito, diferente do experimento, as instâncias também foram consideradas. Neste caso, além do rótulo (*rdfs:label)* e o ponteiro da instância, também é recuperado o conceito ao qual a mesma está associada. O usuário seleciona quais dos recursos apresentados pelo sistema serão utilizados para realizar a expansão da consulta.

A primeira consulta executada realiza a busca nas anotações semânticas usando uma lista de ponteiros para recuperar todas as séries anotadas com estes conceitos, seus conceitos filhos e instâncias de conceitos. A geração da consulta com a lista de palavras semanticamente expandida utiliza a opção de combinação de termos (TODOS ou QUALQUER UM) para acrescentar os rótulos dos recursos selecionados pelo usuário. Esta segunda consulta recupera outro conjunto de séries históricas com base no conteúdo dos metadados descritivos e que não tem interseção com o conjunto resultado da primeira consulta.

A especificação da prioridade em relação aos critérios de qualidade é utilizada para ordenação dos dois conjuntos. Os resultados são apresentados separadamente ao usuário, através de uma lista com os metadados descritivos de cada série, até o limite máximo de séries especificado, conforme apresentado na figura 6.4. Para cada série apresentada, a aplicação disponibiliza o detalhamento dos metadados, o acesso aos dados quando estes estão no estágio publicado, a marcação de relevância na

busca, o detalhamento da anotação semântica, quando existe, e a exploração da ontologia de domínio a partir do recurso usado para anotação.

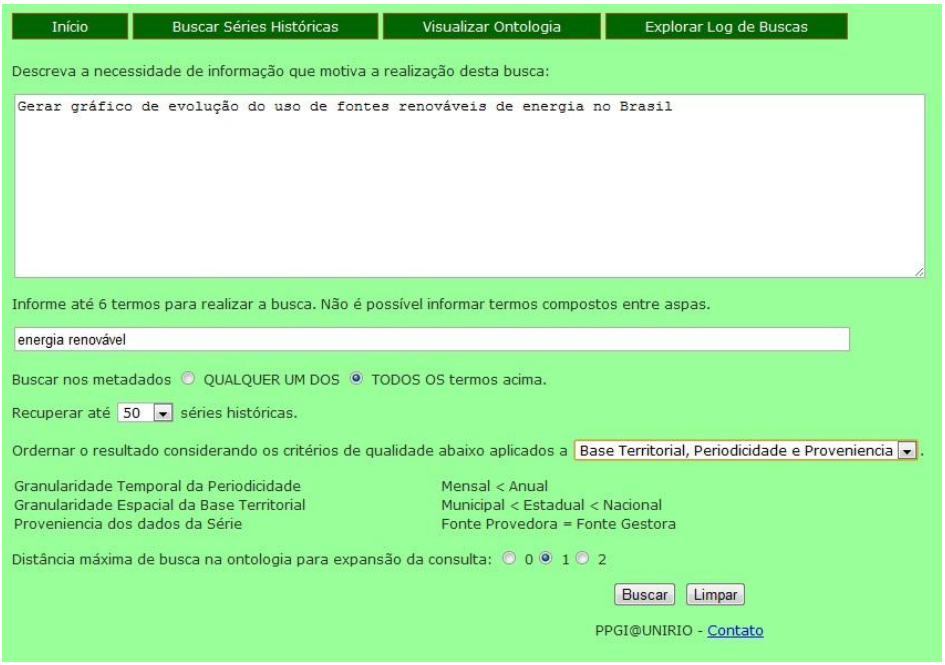


Figura 6.3 Buscar Séries Históricas – prova de conceito

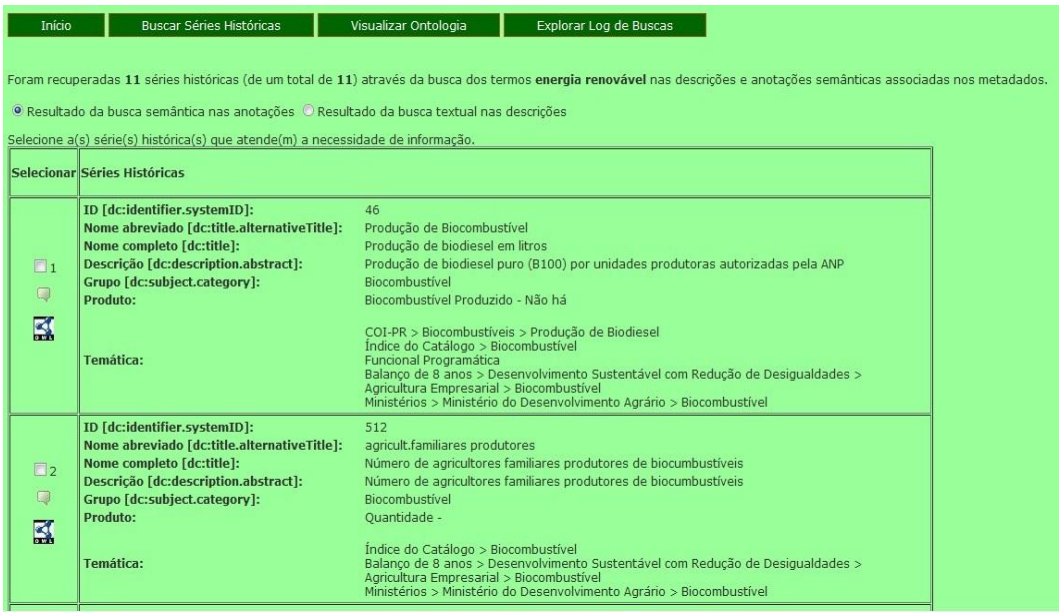


Figura 6.4 Resultado da busca por séries históricas

Todos os parâmetros de busca (descrição da necessidade de informação, palavras-chave iniciais, prioridade dos critérios de qualidade, tipo de combinação das palavras, palavras-chave e ponteiros dos recursos adicionados) e julgamentos de relevância por parte do usuário foram registrados em um *log*, associados a uma identificação única de sessão, para a análise quantitativa e qualitativa. Em uma

mesma sessão, o usuário pode realizar várias buscas para a mesma necessidade de informação com diferentes parâmetros. Através da opção "**Explorar Log de Buscas**" é possível consultar alguns indicadores quantitativos das consultas como, por exemplo, as consultas mais populares e menos populares.

Na próxima seção, é realizada a análise de uma busca por séries históricas utilizando a aplicação da prova de conceito através de um exemplo.

6.4.4 Análise da Busca através de um Exemplo

O fornecimento de energia está em constante mutação em função de alterações na legislação, nas políticas sociais e econômicas internas e no cenário da economia mundial. Além disso, existe uma pressão da sociedade em promover o uso de fontes renováveis para reduzir os efeitos da degradação ambiental. Supondo que um economista necessite mapear a evolução da oferta de fontes de energia no Brasil nos últimos anos e para atender a sua necessidade de informação utilize as séries históricas sobre os indicadores de fontes de energia publicadas no portal *DadosGov* COI-PR, as seguintes atividades deveriam ser realizadas: (1) Buscar as fontes de informação no portal; (2) Recuperar os dados das séries históricas; (3) Integrar os dados e (4) Gerar um gráfico. A aplicação da prova de conceito seria usada para a busca usando palavras-chave no Catálogo de Informações (repositório de metadados) com apoio do VCGE (ontologia de domínio) para expansão da lista de palavras-chave e utilização da associação entre os conceitos e instâncias do VCGE em relação as séries históricas catalogadas (anotação semântica). Este usuário inicia a busca utilizando os termos "***energia***" e "***renovável***", o operador E (AND) e a distância semântica máxima para propagação da busca na ontologia como 1. A escolha da importância das características de qualidade para ordenação dos resultados neste exemplo foi granularidade espacial, granularidade temporal e proveniência.

A partir da ontologia de domínio, o sistema recupera seis conceitos usando a palavra "***energia***" e dois a partir da palavra "***renovável***". Os dois primeiros conceitos recuperados foram **#Energia_Renovavel** e **#Energia_Nao_Renovavel** nesta ordem e estes são conceitos disjuntos e filhos do conceito **#Energia_e_Meio_Ambiente**. Nesta busca somente o conceito **#Energia_Renovavel** foi selecionado para realizar a reformulação da consulta. A figura 6.5 apresenta o conjunto de recursos recuperados pelo sistema e sugeridos para expansão dos resultados enquanto que a figura 6.6 contém o fragmento da ontologia de domínio que representa este mesmo conjunto de recursos recuperados.

As duas consultas (buscas por anotações com a lista de URIs e busca textual pela lista de palavras-chave semanticamente expandida) são executadas e

recuperaram onze séries históricas, nove delas através de anotação e somente duas com a busca textual nos metadados descritivos. As três primeiras séries recuperadas através de anotações correspondiam ao conceito **#BioCombustivel**, que é conceito filho de **#BioMassa** e este por sua vez é filho de **#Energia_Renovavel**. A granularidade espacial dos valores destas séries por Estado justifica a posição destas como as primeiras do resultado. As próximas seis séries, cuja granularidade espacial dos dados é o território nacional, estavam anotadas com os conceitos **#Hidrelétrica** e **#Carvao_Vegetal**, que são conceitos filhos de **#Energia_Renovavel**, e **#Alcool**, que também é um conceito filho de **#BioCombustivel**. As duas séries do conjunto de resultado da busca textual também são relevantes para acompanhar a evolução do uso de fontes renováveis, mas estas não estavam associadas a nenhum tipo específico de fonte; já a descrição destas séries informava somente que contêm dados de "*outras fontes de energia renovável*". Apesar de todas as séries recuperadas estarem relacionadas com o conceito **#Energia_Renovavel**, somente oito delas foram consideradas relevantes por isso seus dados foram recuperados do portal.

Início | Buscar Séries Históricas | Visualizar Ontologia | Explorar Log de Buscas

Resultados retornados pela busca dos termos **energia renovável** na ontologia.

Selecione o(s) tipo(s) de expansão e o(s) termo(s) que serão acrescentados para realizar a busca no repositório de metadados

Termo	Rótulos	Tipo de Expansão
energia	☑ Energias renováveis|	Conceito
	☐ .Biomassa|	.Subconceito
	☐ .Eólica|	.Subconceito
	☐ .Geotérmica|	.Subconceito
	☐ .Hídrica|Hidráulica|Hidrelétrica|	.Subconceito
	☐ .Maremotriz|	.Subconceito
	☐ .Solar|	.Subconceito
	☐ .Energia e meio ambiente|	.Superconceito
	☐ Energia não renováveis|	Conceito
	☐ .Carvão mineral|	.Subconceito
	☐ .Energia nuclear|Energia Radioativa|	.Subconceito
	☐ .Gás natural|	.Subconceito
	☐ .Petróleo|	.Subconceito
	☐ .Energia e meio ambiente|	.Superconceito
	☐ Conservação de energia|	Conceito
	☐ .Energia e meio ambiente|	.Superconceito
	☐ .política e gestão industrial|	.Superconceito
	☐ Energia elétrica|	Conceito
	☐ .Sistema de transmissão de Energia Elétrica|	.Subconceito
	☐ .recursos energéticos|	.Superconceito
	☐ Energia nuclear|Energia Radioativa|	Conceito
	☐ .Urânio|	.Subconceito

Figura 6.5 Resultado da busca por recursos na ontologia

A comparação com as fontes de energia não renováveis também foi contemplada para atender a necessidade de informação que motivou a busca e por isso uma nova busca foi realizada com este objetivo utilizando as palavras "***energia***" e "***não renovável***" e os mesmos parâmetros da anterior. O sistema recuperou da ontologia os mesmos recursos uma vez que a palavra "***não***" foi removida por ser considerada uma *stopword*, mas desta vez somente o conceito **#Energia_Nao_Renovavel** foi selecionado para realizar a reformulação da consulta.

Doze séries foram recuperadas do repositório de metadados, dez por anotação semântica e duas pela descrição dos metadados. As duas primeiras séries, cuja granularidade espacial é por Estado, estavam associadas a ambos os conceitos **#Petroleo** e **#Gas_Natural**, que são filhos de **#Energia_Nao_Renovavel**. As oito séries seguintes, estavam associadas aos conceitos **#Petroleo** (duas), **#Gas_Natural** (duas), **#Carvao_Mineral** (duas) e **#Uranio** (duas). **#Energia_Nao_Renovavel** é o conceito pai (ou superconceito) de **#Carvao_Mineral** e também de **#Energia_Nuclear**, que por sua vez é pai de **#Uranio**. As duas últimas séries recuperadas, do conjunto resposta da consulta textual, são as mesmas da busca anterior pois a palavra "***não***" também foi removida. Mas desta vez estas séries não estavam relacionadas ao conceito **#Energia_Nao_Renovavel**.

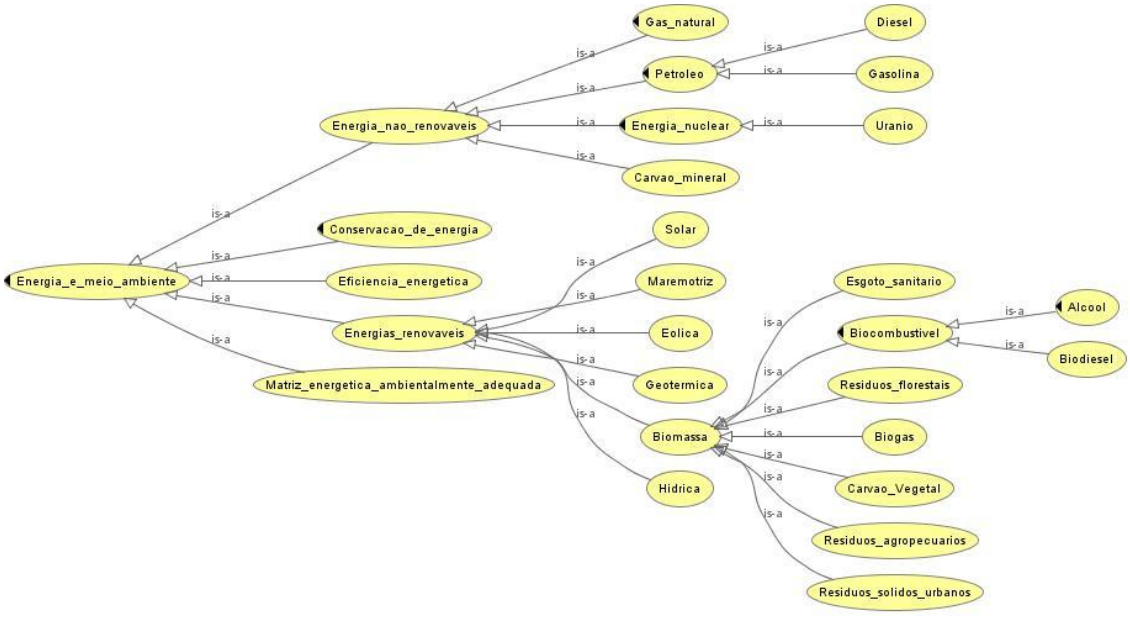


Figura 6.6 Fragmento do VCGE referente a Recursos Energéticos

O cálculo da precisão e cobertura, neste caso, considerada as duas buscas realizadas que recuperaram vinte e três series na mesma seção em relação à necessidade de informação atendida. Neste cenário, o sistema atingiu a precisão de

0,69 (16 séries relevantes recuperadas / 23 séries recuperadas) e cobertura de 1 (16 séries relevantes recuperadas / 16 séries recuperadas).

Oito séries que continham dados sobre o percentual de distribuição de cada fonte foram usadas para confirmar se os dados das outras oito séries correspondiam quantitativamente a 100% do fornecimento de energia no Brasil. Os dados foram integrados para a criação dos gráficos de evolução da matriz energética da figura 6.7, sendo que o agrupamento das séries foi realizado com base na classificação de fontes renováveis e não renováveis do VCGE, conforme apresentada na figura 6.6.

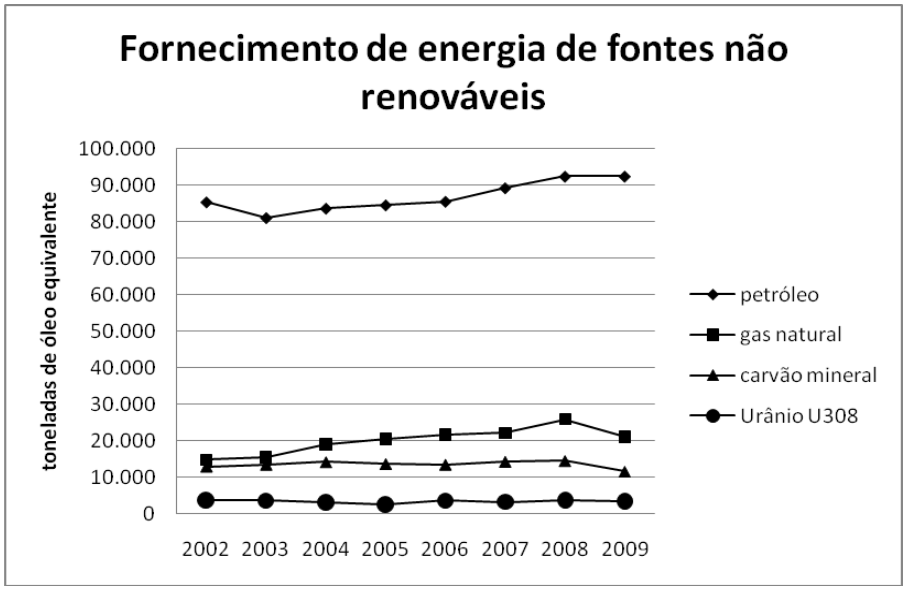
A geração dos gráficos também foi possível, pois todas as séries históricas do grupo de informação “Matriz Energética” possuem a mesma unidade de medida em toneladas equivalentes de petróleo (total de energia gasto pela queima de uma tonelada de óleo cru) e a mesma cobertura temporal de 2002 até 2009 e estas informações estão associadas aos dados através dos elementos Unidade Medida, Início Série e Final Série do esquema de metadados (tabela 6.1).

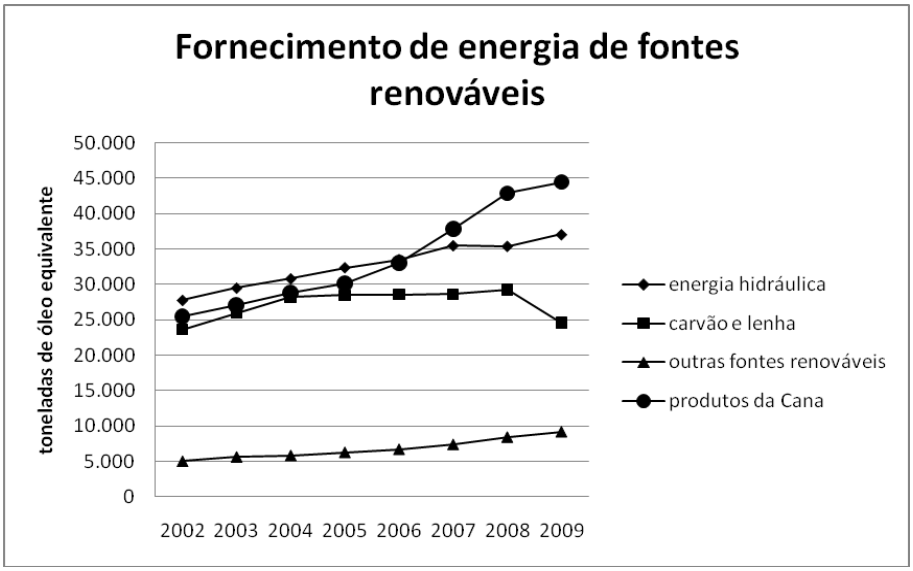
A busca convencional, usando as palavras-chave informadas em cada interação, recuperaria somente as duas series históricas que contêm dados de “outras fontes de energia renovável” e com isso não seria possível gerar o gráfico de evolução da matriz energética com somente duas buscas. O usuário teria que realizar buscas sucessivas utilizando outras palavras (como os rótulos dos conceitos com as quais as series foram anotadas) até recuperar todas as séries e também recorrer a outras fontes para classificar o que é “***Energia Renovável***” ou “***Energia Não Renovável***”. A ontologia de domínio tornou explícita esta classificação das séries permitindo o agrupamento correto dos dados.

6.5 Resultados das Buscas Realizadas com a Aplicação

Após a disponibilização da aplicação na internet [4], um link foi incluído pela equipe do SERPRO na área restrita para manutenção de metadados do portal *DadosGov.* Para participação no estudo de caso foram identificados três perfis de usuários em potencial de dados governamentais federais: (1) organizações civis com foco em transparência na gestão pública ou governo eletrônico, (2) órgãos governamentais que não estão envolvidos na catalogação das séries históricas do portal e (3) cidadãos em geral. Um convite para participação no estudo de caso, com um guia de utilização da aplicação, foi enviado a alguns grupos selecionados que foram contatados por correio eletrônico.

[4] http://semsii.uniriotec.br/semcoi/





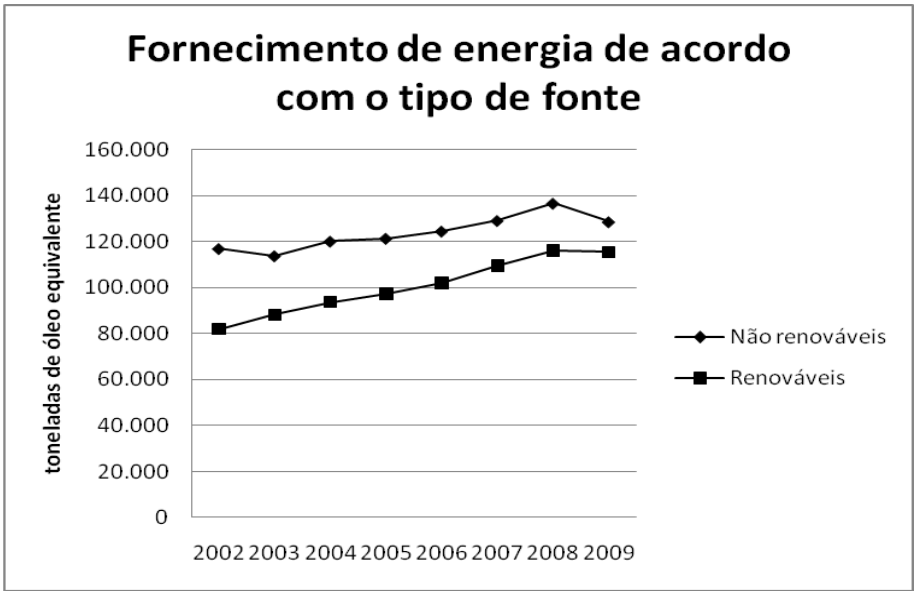


Figura 6.7 Acompanhamento da evolução das fontes de energia

A comunidade Transparência Hacker foi contatada através do seu endereço de correspondência no *GoogleGroups* (thackday@googlegroups.com) por ser um grupo interessado primariamente em dados governamentais abertos. O Observatório de Práticas de Tecnologia da Informação e Comunicação na Gestão Pública (contato@observegov.com.br) foi o segundo grupo da sociedade civil contatado em função de seu objetivo e depois foi realizado um contato com a ONG Contas Abertas (ca@contasabertas.com) por seu interesse em acompanhar, analisar e promover o aprimoramento do dispêndio público. O quarto contato em relação a organizações civis foi feito com o Instituto Acende Brasil (contato@acendebrasil.com.br). Este grupo se define como "*um Centro de Estudos voltado ao desenvolvimento de ações e projetos para aumentar o grau de Transparência e Sustentabilidade do Setor Elétrico Brasileiro*".

Três órgãos governamentais foram contatados. O Instituto de Pesquisa Econômica Aplicada (Ipea), através dos endereços eletrônicos faleconosco@ipea.gov.br e ipeadata@ipea.gov.br, foi selecionado por ser uma fundação pública federal com objetivo de fornecer suporte técnico e institucional às ações governamentais para a formulação e reformulação de políticas públicas e programas de desenvolvimento brasileiros. A Fundação Centro Estadual de Estatística, Pesquisa e Formação de Servidores Públicos do Rio de Janeiro (CEPERJ) foi contatada através dos e-mails egpp@ceperj.rj.gov.br e ceep@ceperj.rj.gov.br. E a terceira instituição deste perfil contatada foi o Instituto Pereira Passos, vinculado à Prefeitura Municipal do Rio de Janeiro e responsável pela aplicação Armazém de Dados, através do endereço de contato armazem@pcrj.rj.gov.br.

Um convite também foi enviado para a lista de alunos do Programa de Pós Graduação em Informática (PPGI) e a lista de pesquisadores do Núcleo de Pesquisa e Prática em Tecnologia (NP^2TEC) de modo a contemplar participantes que representassem o grupo de cidadãos.

No período de 07 de fevereiro de 2011 a 25 de março de 2011 foram registradas 128 consultas em 34 sessões. Em 21 consultas, a descrição da necessidade de informação não foi informada mas 35 necessidades de informação diferentes foram descritas nas 107 consultas restantes. Destas 35, 6 foram descartadas, pois em uma delas (id=18) os termos usados na busca não estavam relacionados com a descrição e nas outras cinco (id=4,7,13,14,25) a descrição não corresponde à uma motivação de busca real por dados governamentais. Cada uma das 29 necessidades de informação restantes foi considerada como uma unidade de análise para o estudo de caso. A tabela 6.6 contém as descrições e o número de

consultas realizadas por necessidade de informação conforme extraídas do *log* da aplicação.

A abordagem utilizada para o estudo de caso não restringiu os cenários de busca como realizado no experimento e por isso não é possível determinar o total de séries históricas relevantes que existem no repositório de metadados para cada necessidade de informação. O julgamento de relevância das séries históricas recuperadas pelo sistema foi realizado pelo próprio usuário a partir da sua intenção de busca e da análise dos metadados descritivos das séries.

Tabela 6.6 Necessidades de Informação dos Participantes do Estudo de Caso

Id	Unidade de Análise	Descrição	Número de consultas
1	Sim	“Minha Casa, Minha Vida” beneficiou 277 mil famílias; 28% da meta	5
2	Sim	TCU encontra “irregularidades graves” no Projovem	2
3	Sim	Ações de educação em áreas rurais	6
4	Não	Apenas um teste de navegação nos dados coletados	1
5	Sim	Combate a violência	1
6	Sim	Como foi o crescimento do software livre nos últimos anos?	1
7	Não	como o governo descreve transparência em seu vocabulário	2
8	Sim	Como o governo Lula deu apoio aos pecuaristas?	1
9	Sim	como o governo trabalhou para a limpeza das cidades	1
10	Sim	compras de software pelo governo federal	4
11	Sim	contratação de serviços de rede pelo governo federal	1
12	Sim	contratação para desenvolvimento de sistemas pelo governo federal	1
13	Não	Convite para participação no Estudo de Caso	3
14	Não	Demonstração para técnicos do SERPRO que participarão do CONSEGI : tema principal DADOS ABERTOS	1
15	Sim	desejava encontrar os dados do SAMU	5
16	Sim	Em 4 anos, governo destinou R$ 194 milhões à ajuda humanitária no exterior	12
17	Sim	estatística de acesso à internet no Brasil	6
18	Não	estou fazendo um mestrado e preciso de informações da área social	1
19	Sim	Fonte de energias renováveis	23
20	Sim	Identificar a quantidade de municípios atendidos por ações de provisão habitacional de interesse social.	2
21	Sim	Identificar ações relacionadas à garantia dos direitos do trabalhador	1
22	Sim	Identificar ações relacionadas a parcerias	1
23	Sim	Identificar dinheiro repassado aos municípios para adequação de segurança pública	3
24	Sim	Identificar iniciativas de melhoria da relação empregado empregador	1
25	Sim	Identificar iniciativas de redução de impostos no mercado de informática	1
26	Sim	Identificar quantidade de abonos salariais PIS/PASEP pagos	1
27	Sim	Investimento e auxílio a educação superior	6
28	Sim	promoção de serviços através da web pelo governo federal	5
29	Sim	Quanto o governo gastou com a distribuição de bolsas aos mais pobres	2
30	Sim	Recuperar informações sobre cisternas.	1

31	Sim	Recuperar informações sobre livros no esporte, especificamente sobre autores.	1
32	Sim	Recuperar informações sobre manutenção rodovias.	1
33	Sim	Recuperar quantidade de alunos escritos no ENEM	1
34	Sim	Tratamento de lixo urbano	2
35	Não	Ver como funciona o sistema.	1

Dezoito consultas recuperaram pelo menos uma série relevante, mas considerando que uma necessidade de informação pode ser explorada por mais de uma consulta, o cálculo de precisão e cobertura foi realizado em relação à cada um das 15 necessidade de informação como um todo e não a cada consulta isoladamente por isso cada necessidade de informação é considerada uma unidade de análise no estudo de caso.

6.5.1 Análise da Precisão

A tabela 6.7 apresenta a comparação da precisão do resultado da busca semanticamente expandida assistida pelo usuário em relação à busca convencional com a lista inicial de palavras-chave. Em média, a aplicação atingiu uma precisão 7,28% maior que a busca convencional. Ao analisar as 15 necessidades de informação é possível observar que a expansão causou redução da precisão em uma, melhorou a precisão em seis e não interferiu no resultado de oito necessidades. Em relação a estas últimas oito, o *log* revelou que duas não foram expandidas por decisão do usuário (23 e 28), outras duas por não recuperarem recursos da ontologia (30 e 33) e as demais por não existirem séries históricas anotadas com os recursos selecionados pelo usuário para expansão.

Tabela 6.7 Comparativo da Precisão do Resultados das Buscas

Id	Sem Expansão			Com Expansão			% Diferença
	Total recuperado	Relevante recuperado	Precisão	Total recuperado	Relevante recuperado	Precisão	
3	17	9	0,529412	22	14	0,636364	20,20%
5	18	5	0,277778	18	5	0,277778	0,00%
6	4	2	0,5	6	4	0,666667	33,33%
8	10	2	0,2	10	3	0,3	50%
9	1	1	1	2	2	1	0,00%
20	29	12	0,413793	56	26	0,464286	12,20%
21	17	2	0,117647	20	3	0,15	27,50%
23	21	4	0,190476	32	9	0,28125	47,66%
24	14	2	0,142857	14	2	0,142857	0,00%
28	5	1	0,2	5	1	0,2	0,00%
29	5	1	0,2	5	1	0,2	0,00%
30	6	6	1	6	6	1	0,00%
31	1	1	1	1	1	1	0,00%

32	4	3	0,75	7	5	0,714286	-4,76%
33	2	1	0,5	2	1	0,5	0,00%
	Total 154	**Total 52**	**Média 0,46813**	**Total 206**	**Total 83**	**Média 0,50223**	**Diferença na média 7,28%**

Quatro necessidades de informação e suas consultas serão analisadas detalhadamente para identificar o comportamento do sistema nestes casos.

Na necessidade de informação “Recuperar informações sobre manutenção de rodovias” (id 32) foi observada a redução da precisão. Para esta necessidade de informação o usuário realizou duas consultas: (1) “***manutenção***” E “***rodovia***” e (2) “***rodovia***' E '***pavimentada***”. Os termos ***“manutenção***” e “***pavimentada***” não recuperam nenhum recurso da ontologia, mas o termo “***rodovia***” com a distância semântica igual a 1 (conforme selecionado pelo usuário) recupera três conceitos, dois conceitos filhos e quatro conceitos pais. O usuário selecionou os conceitos **#Rodovia** e **#Rodovia_Federal**, filho de **#Rodovia** e o sistema recuperou sete séries do Grupo de Informação “Rodovias Federais”, apresentadas na tabela 6.8, pois estas haviam sido anotadas com o conceito #Rodovia_Federal. A abordagem de busca convencional recuperaria somente três séries na primeira consulta e uma na segunda e as séries 157 e 158, apesar de relevantes, não seriam recuperadas com somente duas consultas uma vez que os termos “***manutenção***” e “***pavimentada***” não foram utilizados para sua descrição.

Tabela 6.8 Análise do Resultado das Buscas – Redução da Precisão

Grupo Informação “Rodovias Federais”	**Relevante**	**Recuperada na busca**	
		semântica	**convencional**
ID 160 Manutenção - Rodovias Restauradas	Sim	Sim	Sim
ID 161 Manutenção - Rodovias com Reparos Localizados	Sim	Sim	Sim
ID 162 Manutenção - Total de Rodovias Recuperadas	Sim	Sim	Sim
ID 157 Rodovias Duplicadas	Sim	Sim	Não
ID 158 Rodovias Adequadas	Sim	Sim	Não
ID 156 Rodovias Construídas e Pavimentadas	Não	Sim	Sim
ID 159 Total de Rodovias Construídas, Duplicadas e Adequadas	Não	Sim	Não

De acordo com a interpretação do usuário, o sistema recuperou duas séries consideradas como não relevantes. Estas continham dados sobre construção de rodovias e não sobre manutenção de rodovias, e cinco séries relevantes, nas quais

duas destas não continham os termos "***manutenção***" ou "***pavimentada***". O conjunto de séries recuperadas e o julgamento de relevância deste usuário em particular permitem extrair do *log* de consultas dois conceitos: "***Construção de Rodovias***" e "***Manutenção de Rodovias***". Estes conceitos podem ser considerados como filhos de **#Infraestrutura_de_transporte_rodoviario** (que é o conceito pai de **#Rodovia**). Se estes conceitos fossem incluídos na ontologia com os respectivos rótulos, o sistema recuperaria o conceito **#Manutencao_de_Rodovia** na primeira consulta como o primeiro conceito da lista de sugestões para expansão pois contém ambos os termos "***manutenção***" e "***rodovia***".

Na necessidade de informação "Como o governo Lula deu apoio aos pecuaristas?" (id 8) foi observado o maior aumento (50%) na precisão. Para esta necessidade de informação o usuário realizou somente uma consulta usando um único termo "***pecuarista***" e limitou em dez o número de séries a serem recuperadas pelo sistema. O termo "***pecuarista***", depois de submetido ao processo de redução ao radical, tornou-se "***pecuária***". Estes são conceitos relacionados, pois segundo o dicionário *Michaelis* (Weiszflog 2006) pecuarista é "*Quem se dedica à pecuária ou é versado nela*" mas representam conceitos distintos. O conceito "***pecuarista***" não existe na ontologia, por isso o sistema recuperou somente o conceito **#Pecuaria** da ontologia e cinco conceitos filhos. O usuário selecionou o conceito **#Produto_Animal** (filho de **#Pecuaria**) e o sistema adicionou o rótulo deste conceito à consulta, criando a consulta semanticamente expandida "***pecuária*** OU "***produto animal***".

A palavra "***pecuária***" é encontrada em 110 séries associadas ao item "Ministério da Agricultura, **Pecuária** e Abastecimento" da taxonomia temática "Ministérios". Todas as séries possuem a granularidade temporal como Anual e a fonte gestora dos recursos é a mesma fonte provedora dos dados, mas a granularidade espacial de 39 séries é Estadual e as demais são Nacionais. A expressão "***produto animal***" adicionada na consulta está presente em somente dez séries com granularidade Estadual, todas do grupo "Desenvolvimento do setor agropecuário". A adição do rótulo do conceito **#Produto_Animal** na consulta com o operador de disjunção (OU) não altera a quantidade de séries encontradas pois o conjunto de séries recuperadas usando a expressão "***produto animal***" é um subconjunto das séries do "Ministério da Agricultura, **Pecuária** e Abastecimento".

Neste caso em particular, a ordenação dos metadados pelas características de qualidade das séries não interferiu nos resultados, pois o usuário havia limitado a consulta a dez séries e as dez primeiras tem a mesma granularidade espacial. O aumento da precisão ocorreu porque o critério de ordenação secundário é a pontuação de similaridade entre os metadados das séries recuperadas com expansão

e a consulta expandida "***pecuária***" OU "***produto animal***" enquanto que para a consulta convencional é considerada somente a palavra-chave "***pecuária***".

A tabela 6.9 apresenta a comparação dos resultados em relação ao cálculo da pontuação de similaridade entre a consulta e os metadados descritivos das dez primeiras séries recuperadas em cada abordagem e em negrito as séries consideradas relevantes. A série 349 não está presente nas dez primeiras séries da consulta convencional, pois a sua pontuação de similaridade é igual 0.00632289. Nesta abordagem esta série seria recuperada na vigésima posição.

O maior número de séries relevantes foi recuperado na necessidade de informação "Identificar a quantidade de municípios atendidos por ações de provisão habitacional de interesse social" (id 20). O usuário informou os termos "***Provisão***", "***habitacional***", "***interesse***" e "***social***" e a opção QUALQUER TERMOS para a combinação dos mesmos. Em uma consulta convencional, o sistema recupera 29 séries no total e doze destas são consideradas relevantes pelo usuário uma vez que pertencem ao Grupo de Informação "Provisão habitacional de interesse social".

Tabela 6.9 Análise do Resultado das Buscas – Aumento da Precisão

Consulta convencional: "pecuária"		**Consulta semanticamente expandida: "pecuária" OU "produto animal"**	
ID	Pontuação	ID	Pontuação
2057	0.00915021	348	0.0121106
2058	0.00915021	352	0.0121106
2056	0.00878652	353	0.0121106
342	0.00728943	351	0.0119283
204	0.00703305	349	0.0110651
339	0.00692035	350	0.0106387
348	0.00692035	355	0.0106387
352	0.00692035	354	0.0089621
353	0.00692035	356	0.0082017
351	0.00681619	357	0.0082017

O sistema não recuperou recursos da ontologia com os termos "***provisão***" e "***interesse***", o termo "***social***" recuperou vários recursos, porém nenhum foi selecionado pelo usuário, enquanto que o termo "habitacional" recuperou os conceitos **#Financiamento_habitacional** e **#Seguro_habitacional**, além do conceito **#Habitação** que é pai de ambos. Neste caso, o usuário selecionou os conceitos **#Financiamento_habitacional** e **#Habitação** para adicionar a consulta. A busca nas anotações adicionou 26 séries anotadas com a instância **#Minha_Casa_Minha_Vida** do conceito **#Financiamento_Habitacional**, e destas o usuário selecionou quatorze como relevantes. O aumento do número de séries relevantes recuperadas e também da precisão neste caso ocorreram em função das séries adicionais relevantes possuírem anotação semântica.

O quarto exemplo deste grupo são as consultas associadas à necessidade de informação “Ações de educação em áreas rurais” (id 3). Neste caso, o usuário realizou a primeira consulta com a combinação “***educação***” E “***rural***” e a palavra “educação” recuperou o conceito **#Educacao_no_campo** que foi selecionado pelo usuário. O termo rural, de acordo com o dicionário *Michaelis*, (Weiszflog 2006) significa: “*1 Pertencente ou relativo ao campo ou à vida agrícola; campestre. 2 Próprio do campo. 3 Situado no campo. 4 Agrícola, campesino, camponês, rústico.*”, com isso a expressão “educação rural” pode ser considerada sinônimo de “educação no campo”. A analise do significado dos termos utilizados na consulta, da decisão de expansão do usuário e das séries indicadas como relevantes sugere que “***educação rural***” pode ser uma rótulo alternativo para “***educação no campo***”, pois a consulta “***educação rural***” OU “***educação no campo***” recupera quatorze séries relevantes enquanto que a consulta convencional recupera somente nove delas.

A precisão também pode ser calculada em relação ao conjunto total de consultas realizadas através da aplicação que recuperaram fontes de informação relevantes. Este indicador permite avaliar se a arquitetura proposta colabora com o objetivo do portal *DadosGov* em ser uma iniciativa de disseminação de dados governamentais. O sistema recuperou 206 séries históricas no total, sendo 83 consideradas como relevantes pelos participantes (83/206=0,40291) enquanto que uma funcionalidade de busca convencional por palavras-chave usando o mesmo conjunto de consultas recuperaria 154 séries históricas no total mas somente cinqüenta e duas relevantes (52/154=0,33766). Considerando todo o esforço necessário na coleta, catalogação e publicação destes dados, este resultado evidencia o retorno que a arquitetura pode proporcionar quanto ao aumento da habilidade do sistema para para descobrir e recuperar as informações que atendem às necessidades dos usuários assim como do potencial de reuso das séries históricas publicadas.

6.5.2 Análise da Cobertura

O total de séries históricas relevantes para cada necessidade de informação é indeterminado por isso pode ser considerado como igual a ***N***. No melhor caso, onde a cobertura seria igual a 1, o total de documentos relevantes é igual ao total de documentos relevantes recuperados pelo sistema. No pior caso, a cobertura tende a zero quando o total de relevantes for igual ao total de séries históricas cadastradas (1924) menos as séries recuperadas que não foram consideradas relevantes pelo usuário.

A tabela 6.10 apresenta o cálculo da cobertura presumida no pior caso para cada uma das necessidades de informação que recuperaram pelo menos uma série histórica relevante. A cobertura melhorou em oito necessidades de informação e não interferiu o resultado de sete casos. A modificação de consultas com adição de termos e utilização do operador OU (disjunção), conforme utilizada na arquitetura, nunca irá reduzir as fontes de informação relevantes recuperadas em relação ao universo total de relevantes que existam na coleção.

Tabela 6.10 Comparativo do Resultado da Cobertura das Buscas

Id	Sem Expansão			Com Expansão			% Diferença
	Total relevante	Relevante recuperado	Cobertura	Total relevante	Relevante recuperado	Cobertura	
3	1916	9	0,0047	1916	14	0,00731	55,56%
5	1911	5	0,00262	1911	5	0,00262	0,00%
6	1922	2	0,00104	1922	4	0,00208	100,00%
8	1917	2	0,00104	1917	3	0,00156	50,00%
9	1924	1	0,00052	1924	2	0,00104	100,00%
20	1894	12	0,00634	1894	26	0,01373	116,67%
21	1907	2	0,00105	1907	3	0,00157	50,00%
23	1901	4	0,0021	1901	9	0,00473	125,00%
24	1912	2	0,00105	1912	2	0,00105	0,00%
28	1920	1	0,00052	1920	1	0,00052	0,00%
29	1920	1	0,00052	1920	1	0,00052	0,00%
30	1924	6	0,00312	1924	6	0,00312	0,00%
31	1924	1	0,00052	1924	1	0,00052	0,00%
32	1922	3	0,00156	1922	5	0,0026	66,67%
33	1923	1	0,00052	1923	1	0,00052	0,00%
			Média 0,00181			**Média 0,00290**	**Diferença na media 59,82%**

6.5.3 Análise do *Log* de Consultas

O registro de consultas permitiu identificar que quarenta dos noventa e três termos distintos utilizados nas consultas foram mapeados, total ou parcialmente, em conceitos e instâncias presentes no VCGE. Porém, mesmo mapeados em recursos da ontologia, ainda podem indicar a criação de outros conceitos se nenhum dos recursos recuperados for selecionado pelo usuário para expansão da consulta. Um exemplo é o termo "***violência***" para o qual o sistema recuperou da ontologia o conceito **#Violencia**, filho de **#Saude**, e seus conceitos filhos **#Violencia_contra_o_idoso**, **#Violencia_contra_o_menor**, **#Violencia_domestica** e **#Violencia_urbana**, mas o usuário não selecionou nenhum destes conceitos para a expansão da consulta. A

necessidade de informação do usuário foi descrita como: “Combate a violência” (id 3), o que sugere a criação de um novo conceito associado ao conceito **#Seguranca_Publica**.

No que diz respeito aos termos não mapeados em recursos da ontologia, que recuperam séries históricas, foram encontradas algumas siglas de outros programas de governo como o “Projovem”, “Pronera”, “PNAES” e “Reuni”. A análise do *log* de consultas deve ser realizada por um especialista de domínio para identificar os conceitos que estes programas de governo podem ser associados como instâncias. Outro caso de sigla encontrado é o ENEM (Exame Nacional do Ensino Médio). O conceito **#Ingresso_no_ensino_superior** já existe no VCGE e seus conceitos filhos são **#Transferência** e **#Vestibular**. Um especialista do domínio da educação analisando o *log* de consultas pode decidir que o conceito **#ENEM**, com os rótulos “***ENEM***” e “***Exame Nacional do Ensino Médio***”, deve ser incluído no VCGE como filho de **#Ingresso_no_ensino_superior** e conceito disjunto de **#Transferência** e **#Vestibular**. A inclusão deste conceito não melhoraria o resultado da consulta realizada pelo usuário, pois a sigla ENEM foi utiliza na descrição das séries históricas também, mas torna explícita e precisa a distinção entre as formas de ingresso no ensino superior e aumenta a completude do modelo de domínio.

Os algoritmos de *stemming* e lematização não podem ser aplicados em siglas, pois podem alterar o sentido da consulta ou torná-la sem sentido, como no caso da sigla “Reuni” que o sistema transformou em "('reuni' | 'reunir') | 'unir'" ou a sigla “PNAES” que o algoritmo de *stemming* removeu o “s” com o tratamento de plural.

A remoção de *stopwords* (como conjunções e artigos) ajuda no casamento da consulta em relação aos rótulos e metadados como no caso de “***educação no campo***” e “***educação do campo***”, pois não alterou o sentido. Mas podem ser necessárias para o entendimento da intenção de busca do usuário. Por exemplo, na necessidade de informação “Identificar iniciativas de redução de impostos no mercado de informática” (id 25) o usuário informou as palavras-chave como ‘***imposto em informática’***, o sistema formulou a consulta como “**imposto**” E “**informática**” depois da remoção da preposição “***em***” (considerada como *stopword*) e recuperou dois conceitos **#Imposto** e **#Informatica**. Considerando a descrição da necessidade de informação, a expressão “***imposto em informática***” pode ser interpretada como uma especialização do conceito **#Imposto** em relação ao ramo de atividade das empresas ou tipo de produtos de onde ele é arrecadado (informática). Mas se o usuário não descrever a sua necessidade de informação, a consulta “**imposto**” E “**informática**” poderia ser interpretada com a busca por ferramentas de informática que realizem o controle do

pagamento ou da arrecadação de impostos. Esta avaliação só é possível se o *log* de consultas armazenar a lista inicial de palavras-chave conforme informada pelo usuário.

6.5 Considerações Finais sobre o Estudo de Caso

O estudo de caso, realizado em uma aplicação do governo brasileiro, o Catálogo de Informações *DadosGov* COI-PR, demonstrou os benefícios e as dificuldades da utilização desta arquitetura em um cenário real. A melhoria da precisão e cobertura dos resultados em relação ao conjunto de todas as consultas realizadas indica que a arquitetura proposta colabora com o objetivo do portal *DadosGov* de disseminação de indicadores estratégicos das ações governamentais. Mas o julgamento de relevância dos usuários em relação às séries recuperadas como resultado das consultas é essencial para calcular a precisão, estimar a cobertura, analisar e tomar ações a partir das informações do *log* de consultas.

Através da aplicação do estudo de caso, os usuários realizaram o julgamento de relevância em somente quinze das trinta e cinco necessidades de informação descritas, ou seja, em menos de 50%, e isto pode ter acontecido em função de: (1) o catálogo não contém dados relacionados com vinte das necessidades de informação analisadas, (2) nenhuma das séries recuperadas pelo sistema era relevante para a necessidade de informação que motivou a busca apesar de existirem séries sobre o assunto no portal ou (3) usuários não desejaram informar o seu julgamento de relevância de modo explícito. Em relação a este último motivo, se a funcionalidade de busca estivesse integrada à aplicação do portal seria possível extrair o julgamento de relevância de modo implícito se o usuário, após analisar os metadados recuperados, optasse por visualizar o conteúdo ou realizar o *download* dos dados das séries.

A análise qualitativa e sistematizada do *log* de consultas permitirá identificar novas séries históricas para coleta de dados e publicação no portal e também a priorizar a publicação de séries já coletadas com base no volume de buscas realizadas. A catalogação das séries históricas também deverá ser alterada para que sejam selecionados da ontologia os recursos que devem ser associados, através de anotação semântica, às novas séries disponibilizadas no portal. Caso os conceitos ou instâncias não estejam modelados na ontologia de domínio, o especialista do domínio deverá providenciar a inclusão destes e dos respectivos relacionamentos e atributos.

A catalogação das séries no portal não utilizou o VCGE como um vocabulário de referência, porém, a ontologia criada a partir deste aumentou a precisão do resultado das consultas em seis dos sete casos analisados onde o usuário decidiu expandir a consulta com os recursos sugeridos pelo sistema. Isto pode ser atribuído

ao fato de tanto o vocabulário quanto o repositório de metadados pertencerem ao mesmo domínio e terem sido criados por representantes dos órgãos governamentais. Além disto, os conceitos já existentes no VCGE poderão ser usados para identificar novas séries históricas de interesse para divulgação pelo portal, iniciando um processo de coleta de dados junto aos órgãos setoriais.

Consultas com termos que representam nome de países (“Haiti”) e regiões (“sudeste”) também foram encontradas no log. Nestas situações o usuário desejava especificar lugares (*onde?*), além de assuntos (*o quê?*), relacionados com a sua necessidade de informação. Um filtro da Base Territorial do Grupo de Informações já está disponível na aplicação de visualização de dados do portal, então a integração da funcionalidade de busca na aplicação permitirá que a partir da busca por palavras-chave relacionadas ao assunto, o usuário encontre as séries e depois filtre e agrupe os dados de acordo com a dimensão espacial de interesse.

O próximo capítulo conclui a dissertação, consolidando os resultados obtidos em ambos os métodos de avaliação, e ressalta as principais contribuições e limitações da pesquisa, além de sugerir trabalhos futuros.

7. Conclusão

A atividade de busca por fontes de informação depende do conhecimento do domínio e sua eficiência está relacionada com a completude e a capacidade de exploração do repositório de metadados onde as mesmas estão catalogadas. A utilização de busca semântica torna esta atividade mais eficiente ao utilizar ontologias de domínio que suportam a busca e catalogação destas fontes no repositório de metadados.

A presente pesquisa apresentou uma arquitetura lógica suportada por busca semântica para recuperação de fontes de informação em repositórios de metadados. Nesta arquitetura, ontologias de domínio são utilizadas para formalizar as relações semânticas entre os conceitos de um domínio particular, para recuperar o melhor conjunto de fontes de informação semanticamente similares e aumentar a semântica de metadados descritivos através de anotações semânticas.

A abordagem utilizada identifica a semântica da consulta para direcioná-la às fontes relevantes desta coleção, além de tornar explícita a semântica destas fontes e a similaridade semântica entre elas. A busca semântica permite minimizar os problemas decorrentes de divergência terminológica, conflitos das fontes de informação e perda semântica existentes no repositório de metadados. A expansão semântica da consulta assistida pelo usuário explora o conhecimento do domínio e permite que o usuário especifique os conceitos envolvidos em sua necessidade de informação de uma forma mais explícita, precisa e baseada em sua intenção de busca.

Mas para garantir que todas as fontes de informação relevantes que suportam os processos de negócio, sejam elas internas ou externas a organização, estejam registradas no repositório de metadados, é necessário que políticas e procedimentos organizacionais para registro destas sejam definidos, seguidos e monitorados.

7.1 Considerações Gerais sobre a Avaliação da Proposta

As etapas de avaliação realizadas evidenciaram, através da análise quantitativa de seus resultados, que a arquitetura proposta melhora a eficiência da

busca por fontes de informação no que diz respeito à precisão e cobertura em comparação com busca convencional por palavras-chave. O resultado desta análise em relação aos métodos de avaliação é apresentado na tabela 7.1.

A anotação semântica, realizada em conjunto com a catalogação das novas fontes de informação, permite identificar em quantas e quais fontes de informação os dados de um determinado conceito estão contidos e quais os conceitos presentes em uma fonte de informação específica. Esta característica atende à integração semântica de informação, pois através da ontologia de domínio é possível determinar a similaridade ou relacionamento semântico entre os conceitos associados às fontes de informação recuperadas. O estudo de caso permitiu demonstrar este diferencial da arquitetura ao explorar um exemplo de integração de informações de diferentes séries históricas que foram anotadas manualmente com conceitos relacionados a diferentes tipos de fontes de energia renováveis e não renováveis.

Tabela 7.1 Resultado da Análise Quantitativa segundo os Métodos de Pesquisa

	Precisão		**Cobertura**	
Fase de Avaliação	**Busca convencional**	**Busca semântica**	**Busca convencional**	**Busca semântica**
Experimento	0,2700	0,3620	0,4898	0,5656
	% Aumento	34,07%		15,48%
Estudo de Caso	0,4681	0,5022	0,00181	0,0029
	% Aumento	7,28%		59,82%

Durante o experimento foi possível identificar que, apesar de os usuários terem sido orientados a informar termos que correspondem somente aos conceitos associados às necessidades de informações, alguns termos utilizados nas buscas foram mapeados em outros tipos de recursos. Esta situação sugere que a perspectiva de quem realiza a busca quanto ao que vem a ser um conceito, relacionamento, atributo ou instância nem sempre corresponde a perspectiva de quem modelou a ontologia. Neste sentido, a arquitetura compensa esta divergência ao utilizar rótulos de qualquer tipo de recurso modelado nas ontologias de domínio para realizar o casamento dos termos das consultas. A utilização de outros esquemas de classificação, como glossários, anéis de sinônimos, taxonomias e tesauros, não permitiria o mapeamento destes termos.

A análise qualitativa do questionário revelou que a integração de informações, em algumas empresas, ainda requer que o usuário localize, realize o acesso e integre informações manualmente e que as práticas de governança da informação ainda

precisam atingir níveis mais altos de maturidade para garantir a completeza do repositório de metadados.O rastreamento das consultas e a análise qualitativa do *log* fornecem insumos para evolução da ontologia de domínio e para anotação semântica das fontes que se encontram catalogadas no repositório de metadados. As palavras-chave informadas pelos usuários que realizam as consultas devem ser analisadas por um especialista de domínio, através de um processo de engenharia de ontologias. Com este processo, também é possível agregar novos termos que representam siglas, neologismos, jargões e abreviações particulares do domínio como rótulos alternativos para os recursos existentes, acompanhando a evolução do vocabulário do domínio.

A tabela 7.2 apresenta algumas ações tomadas a partir da avaliação do *log* de consultas quando um termo é mapeado total ou parcialmente em recursos da ontologia ou recupera fontes de informação relevantes. Estas ações visam aumentar a completude das ontologias de domínio e do repositório de metadados, além de aumentar a precisão da descrição das fontes de informação.

Tabela 7.2 Ações Tomadas a partir da Análise do *Log* de Consultas

Perguntas	**Respostas**					
Termo mapeado parcial ou totalmente em recursos da ontologia?	Sim	Sim	Sim	Sim	Sim	Não
Recurso da ontologia selecionado pelo usuário?	Sim	Sim	Sim	Não	Não	-
Recurso da ontologia associado a fontes de informação relevantes?	Sim	Não	Não	Não	Não	-
Termo recupera metadados de fontes de informação relevantes?	-	Sim	Não	Não	Sim	Sim
Ação: Nenhuma. As fontes de informação relevantes já estão anotadas.	X					
Ação: Analisar termo para identificar recurso ou rótulo a ser adicionado na ontologia. **Quem:** Especialistas de domínio				X	X	X
Ação: Analisar fontes relevantes para anotação com os recursos selecionados. **Quem:** Responsável pela fonte de informação.		X				
Ação: Analisar fontes relevantes para anotação com os recursos adicionados. **Quem:** Responsável pela fonte de informação.					X	X
Ação: Identificar fontes de informação não catalogadas no repositório de metadados. **Quem:** Gestor do Repositório de Metadados			X			

A evolução da ontologia é necessária para reduzir a distância entre o vocabulário de quem busca e de quem cataloga as fontes de informação, de modo a torná-la mais completa e melhorar o resultado das consultas ao repositório de metadados. Quando não for possível atingir um consenso e a utilizar uma única ontologia para atender as diferentes perspectivas do domínio, devem ser identificadas

as similaridades semânticas entre as ontologias de domínio para geração dos alinhamentos entre elas (Maedche e Staab 2002).

Além disto, a análise qualitativa do *log* permite descobrir os recursos que foram selecionados com mais freqüência durante as buscas e este indicador pode ser utilizado para priorizar a anotação de fontes de informação legadas, uma vez que este ainda é um processo semi-automático e também para identificar fontes de dados que não estão catalogadas no repositório de metadados.

O estudo de caso demonstrou a viabilidade, benefícios e dificuldades de aplicação da proposta em um ambiente real. Neste tipo de solução de integração de informações, a publicação de novas séries históricas pode ser priorizada com base na demanda da busca por estes dados e o *log* de consultas também fornece indicadores neste sentido.

Repositórios de metadados, como o do portal *DadosGov* COI-PR e de outras iniciativas de disseminação de dados, ainda são destinados principalmente ao uso humano mas à medida que a formalização semântica das ontologias e identificadores uniformes e persistentes são adicionados, estes se tornarão interpretáveis por máquinas também permitindo automatizar algumas atividades do processo de integração de informações.

7.2 Trabalhos Relacionados

Esta seção apresenta um conjunto limitado de trabalhos relacionados. Estes trabalhos foram selecionados por estarem associados ao mesmo tema de pesquisa e com o uso de semântica em integração de informações, recuperação de informação e busca por fontes de informação.

7.2.1 Casamento de Esquemas

Brauner *et. al.* (2008) apresentam uma proposta de construção adaptativa de esquema mediado usado pelo mediador de um sistema de integração de informações. A construção do esquema mediado é realizada através do casamento de esquemas baseado em instâncias recuperadas a partir de chamadas a serviços web e considera que atributos com domínios similares são semanticamente equivalentes.

Um operador, que explora a semelhança semântica do nome dos elementos e métodos de segmentação de palavras com base em um corpus para realizar o casamento entre os esquemas, foi proposto e testado por Islam *et. al.* (2008). O nome dos elementos de dois esquemas são submetidos ao método de segmentação e a identificação da similaridade semântica é calculada usando uma pontuação relativa a

semelhança de listas ordenadas de palavras vizinhas a duas palavras-alvo. Um valor de similaridade mínimo, que pode ser ajustado, é utilizado para gerar os mapeamentos candidatos entre os elementos.

Uma parte da semântica das fontes de informação está em sua descrição, estrutura e instâncias (Rahm e Bernstein 2001), por isso abordagens de casamento de esquemas fazem uso destes insumos, de modo isolado ou combinado, para geração dos mapeamentos. Mas propostas de casamento de esquemas partem da premissa que as fontes de informação a serem utilizadas nesta etapa já são conhecidas, assim a busca por estas fontes em ambientes é uma atividade anterior ao casamento de esquemas.

A arquitetura proposta neste trabalho complementa as abordagens de casamento de esquemas *a posteriori,* pois utiliza o conhecimento do domínio para localizar os candidatos potenciais para esta operação, principalmente em ambientes de grandes organizações onde os sistemas de informação são muitos, sobrepostos e heterogêneos. Desta forma é possível podar candidatos incorretos o mais cedo possível, manter o tempo de execução aceitável destas ferramentas, reduzir o esforço humano para avaliar os mapeamentos gerados entre os esquemas e, principalmente, garantir que somente dados relacionados com o mesmo conceito ou com conceitos semanticamente similares serão combinados.

7.2.2 SRI Semântico

Na literatura existem diversas ferramentas de recuperação de informação que utilizam semântica para melhorar a cobertura e a precisão dos resultados. A análise de algumas abordagens que tratam somente de recuperação de documentos realizada por Mangold (2007) permitiu a elaboração do esquema de classificação que também foi utilizado para descrever a arquitetura proposta. Existem propostas de busca semânticas para outros tipos de bases de dados como *SemSearch* (Lei *et. al.* 2006), (Uren *et. al.* 2008) para busca em uma base de dados RDF e *SHIRI-Querying* (Mrabet *et. al.* 2010) aplicada especificamente em documentos semi-estruturados.

SemSearch está dividida em camadas com funcionalidades específicas. A ferramenta é composta por 5 camadas: *Google-like User Interface, Text Search, Semantic Query, Formal Query Language* e *Semantic Data*. Além de operadores para conjunção (E) e disjunção (OU) dos termos, a ferramenta também utiliza o operador **:** (dois pontos) na forma “assunto: palavra chave”. Por exemplo, a consulta **cidade: Rio de Janeiro** recupera documentos anotados sobre a **instância Rio de Janeiro** da **classe cidade** e não recupera documentos anotados sobre a **instância Rio de Janeiro** da **classe estado**. Um processo de tradução de consultas, realizado nas

camadas *Text Search, Semantic Query* e *Formal Query Language,* recebe palavras-chave como entrada e gera consultas em linguagem formal que são submetidas para a camada de dados (*Semantic Data*). Cada palavra é mapeada em conceitos, relações entre conceitos ou até mesmo instâncias utilizando os rótulos (*rdfs:label*) para conceitos e relações e os literais das propriedades para as instâncias. O resultado desta busca são as triplas RDF e também documentos HTML recuperados através de anotação.

A anotação semântica dos documentos semi-estruturados em *SHIRI-Querying* é realizada no nível de granularidade dos nós dos documentos (*tags* XML e HTML) e não no nível dos termos presentes no conteúdo do documento. Cada nó pode ser anotado como contendo instâncias de um (S*etOfConcepts*) ou vários conceitos diferentes (*PartOfSpeech*) de uma determinada ontologia. O critério de ordenação dos resultados recuperados estabelece que os nós que contêm somente instâncias do mesmo conceito têm mais relevância que os nós que contêm instâncias de conceitos diferentes, mesmo que estes conceitos estejam semanticamente relacionados.

O principal diferencial da arquitetura em relação às ferramentas de busca semântica de propósito geral em repositório de documentos é a definição do critério de ordenação do resultado estar vinculado a particularidades da tarefa que a ferramenta suporta. A qualidade das fontes de informação é um fator importante para o processo de integração de informações e depende do contexto da necessidade de informação que motivou a busca.

O componente de reformulação semântica de consulta da arquitetura proposta, assim como *SemSearch*, também realiza uma busca textual com termos da consulta nos rótulos dos recursos da ontologia, que podem ser conceitos, relacionamentos, atributos e até mesmo instâncias. Os esquemas de metadados definem a estrutura para descrição das fontes de informação no repositório de metadados. Assim, os metadados podem ser tratados como documentos semi-estruturados, mas diferente de *SHIRI-Querying,* a anotação das fontes de informação não é realizada no nível de cada elemento do esquema de metadados uma vez que o objetivo da anotação semântica é tornar mais explícita e precisa a descrição de cada fonte de informação.

7.2.3 Extensão de Sistemas de Registro de Metadados

Py *et. al.* (2009) propõem um método com passos detalhados para identificação de conceitos e fontes de dados durante um processo de integração de dados. O método se baseia na análise de glossários de modelos de processos de negócio e assume que todos os processos possuem uma única identificação e descrição para uma entidade, o que pode não acontecer em casos reais.

O método é suportado por um framework que utiliza serviços dependentes de domínio (serviços de dados e serviços de conceitos) e independentes de domínio (serviço de metadados e serviços de integração) além de um repositório de metadados onde as fontes de dados permanentes e os conceitos do negócio, identificados durante a análise dos glossários, e os serviços dependentes de domínio, que foram construídos para suportar a integração de informações, devem ser catalogados e relacionados. Os serviços de dados encapsulam as particularidades de acesso, realizam consultas sobre as fontes de dados e são invocados somente por serviços de conceitos. Os serviços de conceito por sua vez fornecem uma interface comum e unificada para acesso a um conjunto de entidades concretas que podem estar espalhadas em diferentes fontes, mas possuem a mesma semântica definida no escopo do negócio. Os serviços de metadados são responsáveis pelo acesso e recuperação de dados do repositório de metadados. Os serviços de integração são subdivididos em serviços de unificação e serviços e serviços de reestruturação.

Mas somente expor as fontes de dados como serviços não é o suficiente para garantir que a integração semântica das informações atenda aos requisitos de flexibilidade, extensibilidade e facilidade de evolução associados às iniciativas de Arquitetura de Informações (Patrick 2005). Neste sentindo, a arquitetura proposta por este trabalho estende o framework adicionando funcionalidades de busca e anotação das fontes de informação catalogadas no repositório de metadados. Isto torna explícita e formal a semântica destas fontes em relação a um modelo de conceitos, e não requer a definição de um esquema mediado único que represente todas as fontes para a sua integração. Além de tratar problemas de divergência terminológica, conflitos das fontes de informação catalogadas e aumentar a semântica do repositório de metadados.

Dois serviços independentes de domínio devem ser acrescentados ao *framework*: serviço de ontologias de domínio, que atua como o Gerente de Acesso a Ontologias, e o serviço de reformulação de consulta, que desempenha o papel de Reformulador Semântico de Consultas. A interface do serviço de metadados, já disponível no *framework*, também deve ser alterada para que, além da busca pelos nomes dos conceitos e serviços, também seja possível fornecer uma lista de URIrefs para realizar a busca nas anotações semânticas e uma lista de palavras-chave para realizar a busca no conteúdo dos elementos descritivos selecionados do esquema de metadados.

Além dos sistemas de manutenção de repositórios de metadados corporativos, os sistemas de informações geográficas e ambientais e as bibliotecas digitais também são exemplos de aplicações que fazem uso intensivo de metadados para localizar e

integrar informações. Estes sistemas são ferramentas importantes para disseminar e compartilhar conhecimento sobre os recursos disponíveis dentro de uma comunidade e tem sido amplamente utilizados para navegar e recuperar objetos em coleções de imagens, vídeos, áudio, documentos eletrônicos e conjuntos de dados (*datasets*). Nestes sistemas existem problemas para reconciliar os diferentes vocabulários usados pelos usuários para publicar, localizar e interpretar estas descrições. Alguns utilizam paradigmas de catalogação e navegação através de listas de classificação em assuntos ou tesauros (Kashyap *et. al.* 2008). Porém, paradoxalmente, quanto mais objetos são catalogados em repositórios de metadados mais difícil tem sido localizar objetos relevantes. Algumas propostas surgiram para melhorar a eficiência da busca por objetos nestes catálogos.

Um protótipo que estende o sistema de registro de metadados *Metacat* foi apresentado por Berkley *et. al.* (2009). Este sistema é usado por grupos de pesquisa em ecologia relacionados com a Rede de Conhecimento para Biocomplexidade (Knowledge Network for Biocomplexity – KNB). Neste sistema, as fontes de dados são descritas usando o padrão de metadados *Ecological Metadata Language* (EML), que contém vários elementos em texto livre. Nesta versão do sistema são utilizadas ontologias em *OWL-DL* e anotações semânticas para ligar as fontes de dados com conceitos da ontologia. Além da OBOE (*Extensible Observation Ontology*), que é uma ontologia para representar a semântica em um alto nível de abstração de observações e medições científicas, ontologias específicas de domínio descrevem um número limitado de conceitos que são relevantes para uma organização, comunidade ou grupo de pesquisa e são criadas como extensões da OBOE. Três tipos de busca foram adicionados ao sistema: (1) expansão automática da lista de palavras-chave com termos da ontologia, que correspondem a sinônimos e conceitos filhos, (2) busca em anotações através de palavras-chave usando conceitos e conceitos filhos e (3) busca estruturada sobre as anotações através de termos de ontologia. Através de testes com o protótipo, os autores relatam melhoria de precisão e cobertura das novas buscas, se comparado com a busca convencional por palavras-chave.

Da mesma forma, na arquitetura proposta, a consulta por palavra-chave é reformulada adicionando rótulos de recursos da ontologia para busca textual e o URIref dos recursos para a busca usando anotações. Mas, diferente da proposta de Berkley *et. al.* (2009), a busca com propagação por recursos vizinhos prevista na arquitetura proposta não está restrita a conceitos filhos, esta é determinada por um conjunto de regras de propagação que exploram outros relacionamentos semânticos entre os recursos da mesma ontologia de domínio e a similaridade entre recursos de ontologias distintas que foram previamente alinhadas. Estas regras. usam os recursos

das ontologias de domínio mapeados por cada termo da consulta como ponto de partida para a expansão da consulta original. Porém, a arquitetura segue a abordagem de expansão assistida pelo usuário a fim de evitar que a reformulação da consulta torne o resultado distante da intenção de busca do usuário e de permitir que este adquira conhecimento sobre o domínio ao interagir com os fragmentos da ontologia.

7.2.4 Expansão de Consultas Utilizando Ontologias

Na literatura, existem diversos trabalhos que utilizaram a avaliação empírica através de experimentos para analisar o resultado de expansão de consultas utilizando ontologias. Dois exemplos são (Abdelhamid *et. al.* 2009) e (Elias 2010).

Abdelhamid *et. al*. (2009) apresenta um experimento em seis fases para comparar a precisão do resultado de consultas usando o conjunto de documentos *Cystic Fibrosis*, que é um subconjunto de documentos da coleção MEDLINE, anotados com conceitos do MeSH. A primeira fase definiu o *baseline* de comparação e as fases seguintes usaram técnicas variadas para expansão da consulta e ordenação dos documentos. Os autores concluíram que a precisão foi maior quando foi combinada a expansão da anotação semântica com palavras-chave (última abordagem) em função da falta de completude das anotações. Esta abordagem atingiu um aumento de 7.5% na precisão dos primeiros dez documentos.

Elias (2010) avaliou o efeito de cada uma das relações léxico-semânticas presentes em ontologias e outros sistemas terminológicos para a expansão de consultas. O experimento usou a base de documentos TREC Genômica e algumas bases de conhecimento *Gene Ontology*, MeSH, SNOMEDCT, UMLS, MTH por serem específicas do domínio Biomédico e consideradas como ontologias de referência. Cada consulta foi expandida manualmente usando uma relação semântica por vez (sinonímia, hiperonímia, hiponímia, meronímia, holonímia e uma relação específica de domínio) e o resultado foi avaliado comparativamente em relação a um *baseline* (busca convencional). Neste trabalho foi concluído que somente a expansão através da sinonímia melhorou a precisão do resultado da busca (10,31% de aumento para os primeiros 10 documentos) e que a falta de consistência no uso de algumas relações nas ontologias de domínio (como a generalização e especialização) prejudica o resultado da expansão.

Assim como Abdelhamid *et. al*. (2009), a arquitetura proposta combina a busca nas anotações semânticas com a expansão semântica da lista de palavras-chave para compensar a falta de completude das anotações no repositório de metadados, principalmente em relação as fontes de informação que não foram anotadas durante a sua catalogação. Quanto as relações léxico-semânticas exploradas, a arquitetura

utiliza todos os rótulos do mesmo recurso, ou seja, todos os termos considerados como sinônimos para representação de um recurso (conceito, relacionamento, atributo ou instância) para expansão da lista de palavras-chave além de explorar os alinhamentos entre ontologias que identificam recursos equivalentes. Desta forma restringe a relação de sinonímia ao domínio das ontologias selecionadas do repositório de ontologias.

A primeira fase de avaliação da arquitetura proposta contemplou a realização de um experimento. Este experimento comparou a precisão e a cobertura das consultas em relação ao *baseline* de busca convencional (sem expansão) e a expansão estatística (baseada na co-ocorrência de termos), e teve a participação humana na utilização de um protótipo. Uma ontologia de domínio existente foi utilizada e estendida para refletir os conceitos, relacionamentos e atributos presentes nos sistemas de informação a serem integrados. Porém, esta não é uma ontologia de referência, pois a própria arquitetura fornece insumos para a sua evolução a partir dos termos extraídos do *log* de consultas.

7.3 Contribuições

A principal contribuição desta pesquisa é a arquitetura lógica a ser usada como referencial para extensão das funcionalidades de busca e catalogação de fontes de informação em repositório de metadados. A implantação desta extensão não requer que a ontologia de domínio inicial seja completa e nem que todas as fontes de informação catalogadas possuam anotações semânticas associadas, pois a própria arquitetura, através do registro das consultas, fornece insumos para os processos de Engenharia de Ontologias e Anotação Semântica.

Como contribuições secundárias podem ser citados:

- Um resumo dos principais conceitos envolvidos e abordagens adotadas para Integração de Informações, dando destaque à perspectiva semântica e ao papel dos modelos de dados que compõem a Arquitetura de Informações de uma Organização, em especial das Ontologias e Metadados na atividade de descoberta de fontes de informação organizacionais.
- O protótipo SEM-SII construído para a realização do experimento.
- A análise quantitativa e qualitativa dos resultados do experimento e do estudo de caso.
- A análise qualitativa dos aspectos do perfil dos usuários e do ambiente organizacional onde estes estão inseridos relacionados com as

atividades de Integração de Informações a partir.das respostas dos participantes do experimento ao questionário.

- A ontologia de domínio da Administração Pública modelada a partir do Vocabulário Controlado do Governo Eletrônico.
- Aplicação desenvolvida como prova de conceito para realização da busca no Catálogo de Informações do portal *DadosGov* COI-PR.
- A abordagem de expansão semântica de consultas por palavras-chave a partir de regras de propagação aplicadas em ontologias de domínio.
- A abordagem de evolução de ontologias de domínio a partir da análise dos termos extraídos do *log* de consultas da arquitetura.
- A abordagem de anotação semântica das fontes de informação catalogadas a partir da análise dos recursos selecionados e do julgamento de relevância das fontes de informação recuperadas extraídos do *log* de consultas da arquitetura.

7.4 Limitações da Proposta e Trabalhos Futuros

Algumas limitações da proposta merecem ser ressaltadas: (1) a arquitetura não explora o perfil dos usuários e nem o histórico das consultas destes durante a busca por fontes de informação; (2) o aumento da precisão e cobertura do resultado da busca depende da anotação semântica das fontes de informação; (3) o julgamento de relevância das fontes de informação recuperadas é essencial para a avaliação do desempenho da arquitetura e (4) a arquitetura possui como premissa a existência de um repositório centralizado para recuperar os metadados das fontes de informação relevantes.

Vários trabalhos futuros são sugeridos a seguir, para refinar a proposta e lidar com as limitações.

O protótipo da arquitetura utilizou somente a visualização da ontologia através de lista identada. Outras formas de visualização (Katifori *et. al.* 2007) devem ser analisadas para avaliar o impacto na eficiência da busca e anotação de fontes de informação do ponto de vista da usabilidade da ferramenta.

A arquitetura utiliza regras de propagação para busca de recursos na ontologia. Estas regras são definidas para atingir uma distância semântica máxima *d* em relação a cada recurso identificado a partir de cada termo da lista de palavras-chave da consulta inicial do usuário. Novas regras de propagação podem ser definidas e avaliadas como, por exemplo, regras que considerem a distância semântica como a

contagem de arestas do menor caminho entre os recursos (Petrakis *et. al.* 2006) que duas ou mais palavras-chave informadas na consulta recuperam da ontologia.

O registro das consultas pode ser usado como insumo para evolução das ontologias e anotação semântica de fontes legadas. A análise deste *log* foi feita manualmente, mas a utilização destes insumos requer a sistematização (Buitelaar *et. al.* 2005) e automação desta atividade. Mecanismos de mineração de dados poderiam ser utilizados para descobrir conhecimento relevante a partir deste *log* e sugerir ações de melhoria no ambiente, ou mesmo automatizá-las. Este histórico de consultas também poderia ser utilizado no momento da busca para identificar os domínios de interesse e as preferências dos usuários quanto aos critérios de qualidade.

A avaliação da proposta considerou a eficiência da abordagem em relação ao resultado da busca e por isso utilizou as medidas de precisão e cobertura. Esta avaliação depende do julgamento de relevância do resultado realizado de modo explícito pelo usuário. Outras formas de coletar este julgamento de modo implícito também deveriam ser previstas, como por exemplo, a ação do usuário em realizar o acesso ao conteúdo da fonte a partir de seus metadados.

Além de precisão e cobertura, a eficiência também pode ser avaliada em relação ao tempo de resposta da busca por recursos na ontologia e das consultas geradas para recuperação de metadados uma vez que é esperado o aumento da quantidade de recursos nas ontologias, do número de ontologias de domínio utilizadas para anotação das fontes de informação e do número de fontes de informação no repositório de metadados.

A busca em repositórios de metadados distribuídos requer adaptação da arquitetura proposta de *stand-alone* para *meta-search* de modo a distribuir as consultas geradas pelo Reformulador Semântico de Consultas , integrar e ordenar os resultados a serem apresentados.

Os elementos descritivos do esquema de metadados selecionados para a busca textual possuem o mesmo peso. A atribuição de pesos diferentes de acordo com o propósito de cada elemento pode ser considerada para calcular a pontuação de similaridade entre a consulta (da lista semanticamente expandida de palavras-chave) e os metadados, e utilizada como critério de ordenação secundário do resultado da busca nos metadados descritivos.

Avaliação de ferramentas e técnicas de anotação semântica (Reeve e Han 2005) aplicadas a repositórios de metadados para permitir a anotação de fontes de informações legadas assim como sugerir os recursos a serem utilizados para anotação no momento de catalogação de novas fontes.

A expansão de consulta da arquitetura adotou somente a abordagem de busca semântica assistida pelo usuário, pois utilizou as ontologias de domínio para sugerir novos recursos e termos e o conhecimento do usuário que estava realizando a busca para escolher os recursos e termos a serem adicionados a consulta original. Porém, considerando que, durante o experimento, a expansão semântica automática atingiu uma precisão maior entre as abordagens e que a expansão estatística automática poderia usar também o *log* de consultas para extrair padrões de co-ocorrência de termos, a combinação das abordagens de expansão permitiria atingir melhor desempenho no que diz respeito à precisão e cobertura dos resultados. A arquitetura poderia combinar esses e outros métodos de expansão como refinamentos iterativos baseados no julgamento de relevância, com ou sem apoio do usuário que realiza a busca (Manning *et. al.* 2009). Abordagens para expansão de consultas que combinam padrões de co-ocorrência de termos e modelos de conhecimento independentes do corpus já foram avaliadas por outros autores (Schatz *et al.* 1996 *apud* Bhogal *et. al.* 2007, Mandala *et al.* 1999 *apud* Bhogal *et. al.* 2007, Huang *et al.* 2005 *apud* Bhogal *et. al.* 2007) .

Os elementos relativos a qualidade de dados existentes no esquema de metadados são utilizados para a ordenação dos resultados. A arquitetura poderia prever a utilização de modelos de qualidade de dados que permitissem a medição quantitativa de aspectos objetivos e subjetivos referentes da qualidade dos dados destas fontes como o de Simmham (2007) e a associação dos elementos do esquema de metadados aos fatores de qualidade destes modelos.

Metadados também podem ser publicados juntos dos recursos que descrevem ao invés de serem armazenados em um repositório de metadados. A arquitetura poderá futuramente ser adaptada para remover a premissa da existência do repositório de metadados, realizando a busca diretamente nos metadados das fontes de informação disponíveis em rede, e esta mudança permitiria a utilização da arquitetura em buscas na Web quando as fontes forem publicadas com os seus respectivos metadados.

Referências

ABDELHAMID, E., RAFEA, A., EL-BELTAGY, S. R. “Enhancing search results of concept annotated documents”. In: Kang Zhang, Reda Alhajj (eds.), *Proceedings of the 10th IEEE international conference on Information Reuse and Integration (IRI'09).* IEEE Press, pp. 330-335. 2009

AGRAWAL, R., AILAMAKI, A., BERNSTEIN, P.A., *et al.* “The Claremont report on database research”. *Communications of the ACM*, Vol. 52, Nr.6, pp. 56-65, 2009

ALEXIEV, V., BREU, M., BRUIJN, J., *et al.* “Information Integration with Ontologies: experiences from an industrial showcase”. Editora John Wiley & Sons, ISBN 978-0470010488, 2005.

ANSI/NISO “Z39.19-2005 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies”, NISO Press, 176 p., ISBN: 1-880124-65-3, 2005

AZEVEDO, L. A., SIQUEIRA, S. W. M., BAIÃO, F. A., *et al.*: “Enterprise Ontology Management - An Approach based on Information Architecture”. In: *Proceedings of Eleventh International Conference on Enterprise Information Systems* (ICEIS 2009). INSTICC Press, pp. 243–249, 2009

IEEE Computer Society. "IEEE Recommended Practice for Architectural Description of Software-Intensive Systems", *IEEE Std 1471-2000*, 2000

BASS, L., CLEMENTS, P., KAZMAR, R. Software Architecture in Practice. Primeira Edição, Editora Addison-Wesley Professional, ISBN 978-0321154958, 560 p., 1998

BATINI, C., LENZERINI, M., NAVATHE S. B. “A Comparative Analysis of Methodologies for Database Schema Integration”. *ACM Computing Surveys* (CSUR), Vol. 18, Nr. 4, pp. 323-364, 1986

BATINI, C., SCANNAPIECO, M. "Data Quality: Concepts, Methodologies and Techniques: Data-Centric Systems and Applications". Springer-Verlag, ISBN:3540331727, 2006.

BERKLEY, C., BOWERS, S., JONES, M. B., *et al.* “Improving data discovery for metadata repositories through semantic search”. *Third International Conference on Complex, Intelligent and Software Intensive Systems* (CISIS 2009), pp. 1152-1159, 2009.

BERNERS-LEE, T., HENDLER, J., LASSILA, O. “The Semantic Web”, *Scientific American*, pp. 29-37, 2001

BERNSTEIN, P. A., HAAS, L. “Information Integration in the Enterprise”. *Communications of the ACM - Enterprise information integration: and other tools for merging data*, Vol. 51, Nr. 9, pp. 72-79, 2008

BLEIHOLDER, J., NAUMANN, F. “Data Fusion”. *ACM Computing Surveys*, Vol. 41, Nr. 1, Artigo 1, 41 pages, 2008.

BHOGAL, J., MACFARLANE, A., SMITH, P. “A review of ontology based query expansion”.

Information Processing and Management: an International Journal, Vol. 43, Nr. 4, pp.866-886, 2007

BOTTO, R. “Arquitetura Corporativa de Tecnologia da Informação”. Primeira Edição,Editora Brasport, 268 p., ISBN: 8574521779, 2004.

BRAUNER, D. F., GAZOLA, A., CASANOVA, M. A., BREITMAN, K. K. “Adaptative matching of database web services export schemas”. In: J. Cordeiro, J. Filipe (eds), *Proceedings of the Tenth International Conference on Enterprise Information Systems* (ICEIS), Vol. DISI, pp. 49-56, 2008.

BREITMAN, K. K., CASANOVA, M. A., TRUSZKOWSKI, W. “Semantic Web: Concepts, Technologies and Applications”. Springer, Primeira Edição, 330p., ISBN 978-1-84628-581-3 2007.

BUITELAAR, P., CIMIANO, P., MAGNINI, B. “Ontology Learning from Text: An Overview”. In Paul Buitelaar, P., Cimiano, P., Magnini B. (eds.), *Ontology Learning from Text: Methods, Applications and Evaluation,* IOS Press, pp. 3-12, 2005

CASANOVA, M. A., BREITMAN, K. K., BRAUNER, D. F., MARINS, A. L. A. “Database conceptual schema matching”. Computer, *IEEE Computer Society Press*, Vol. 40, Nr. 10, pp. 102-104, 2007.

CASTRO, L., BAIÃO, F. A., GIANCARLO, G. “A Linguistic Approach to Conceptual Modeling with Semantic Types and OntoUML”. In: *5th Joint International Workshop on Vocabularies, Ontologies and Rules for The Enterprise (VORTE) Metamodels, Ontologies and Semantic Technologies*. Workshop Proceedings of the 15th International Enterprise Computing Conference (EDOC 2010), 2010.

CHOI, N., SONG, I., HAN H. “A Survey on Ontology Mapping”. *ACM SIGMOD Record*, Vol. 35, Nr. 3, pp. 34-41, 2006

CONCAR – Comitê de Estruturação de Metadados Geoespaciais do Brasil (CEMG). Perfil de Metadados Geoespaciais do Brasil – Perfil MGB. Dezembro 2009

CRESTANI. F. “Application of spreading Activation Techniques in Information Retrieval”. *Artificial Intelligence Review*, Springer Netherlands, Vol. 11, Nr. 6, pp. 453-482, 1997

CUI, Z., JONES, D., O'BRIEN, P. “Issues in Ontology-based Information Integration”. *International Joint Conference on Artificial Intelligence - Workshop on E-Business and the Intelligent Web* (IJCAI-01), 2001.

DAVENPORT, T. H., PRUSAK, L. “Information Ecology: Mastering the Information & Knowledge Environment”. Editora Oxford University Press, 1997

DCMI (Dublin-Core Metatata Institute). DCMI Metadata Terms, versão de 11/outubro/2010. Disponível em http://www.dublincore.org/documents/dcmi-terms/. Último acesso em 03/janeiro/2011.

DOAN, A., HALEVY, A. Y. “Semantic integration research in the database community: A brief survey”. *AI Magazine*, Vol. 26, Nr. 1, pp. 83-94, 2005.

D. FENSEL. “Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce”. Spring Verlag, 2001.

ELIAS, ANDRÉ BECHARA. *Expansão semântica de consultas baseada em esquemas terminológicos: uma experimentação no domínio biomédico.* UFRJ, 2010. Dissertação de Mestrado. Instituto de Matemática, Núcleo de Computação Eletrônica.

EVERNDEN, R., EVERNDEN, E.: “Third-generation information architecture”. *Communications of the ACM*, Vol. 46, Nr.3, pp. 95-98, 2003.

e-PING. Padrões de Interoperabilidade de Governo Eletrônico. Documento de Referência, Versão 2011, 3 de dezembro de 2010

e-PMG. Padrão de Metadados do Governo Eletrônico. Versão 1, Janeiro de 2010

FERREIRA, R., MOURA-PIRES, J. ”Extensible Metadata Repository for information systems and enterprise applications”. In: *Proceedings of the Ninth International Conference on Enterprise Information Systems* (ICEIS 2007), Volume DISI, pp. 344-350, 2007.

GARLAN, D. “Software Architecture: a Roadmap”. In: *Proceedings of the Conference on The Future of Software Engineering* (ICSE '00). ACM, pp. 91-101, 2000.

GERTZ, M.,¨OZSU, M. T., SAAKE, G., SATTLER, K. “Report on the Dagstuhl Seminar - Data Quality on the Web”. *ACM SIGMOD Record*, Vol. 33, Nr. 1, pp. 127-132, 2004

GIUNCHIGLIA, F., ZAIHRAYEU, I., “Lightweight ontologies”. In: Liu, L., Ozsu, M. T. (eds), *Encyclopedia of Database Systems*, Springer, pp. 1613-1619, 2008.

GODINEZ, M., HECHLER, E., KOENIG, K., *et al.*: The Art of Enterprise Information Architecture - A Systems-Based Approach for Unlocking Business Insight. IBM Press. 2010

GUARINO, N. “Formal Ontology and Information Systems”. In: *Proceedings of the First International Conference on Formal Ontology in Information Systems* (FOIS’98), pp.3-15, 1998.

GUIZZARDI, G. ; BAIÃO, F. ; LOPES, M ; FALBO, R. “The Role of Foundational Ontologies for Domain Ontology Engineering: An Industrial Case Study in the Domain of Oil and Gas Exploration and Production”. *International Journal of Information System Modeling and Design*, v. 1, Vol. 1, Nr. 2, pp. 1-22, 2010.

GRUBER, T. R. “A translation approach to portable ontology specifications”. *Journal Knowledge Acquisition - Special Issue: Current Issues in Knowledge Modeling*. Vol. 5, Nr. 2, pp. 199-220, 1993.

HARRIS, S., GIBBONS, J., DAVIES, J., *et al.* “Semantic Technologies in Electronic Government - Tutorial and Workshop”. In: *Proceedings of the 2nd international conference on Theory and practice of electronic governance* (ICEGOV '08). ACM, pp. 45-51, 2008

HALEVY, A. “Information Integration”. In: Liu, L., Ozsu, M. T. (eds), *Encyclopedia of Database Systems*, Springer, pp. 1490-1496, 2008.

HALEVY, A., RAJARAMAN, A., ORDILLE, J. “Data integration: the teenage years”. In: Umeshwar Dayal, Khu-Yong Whang, et. al. (eds.), *Proceedings of the 32nd international Conference on Very Large Data Bases,* (VLDB '06)*,* 2006.

HOANG, H. H., TJOA, A. M. “The State of the Art of Ontology-based Query Systems: A Comparison of Existing Approaches”. In: *Proceedings of the IEEE International Conference on Computing & Informatics*, 2006.

HOBERMAN, S., BURBANK, D., BRADLEY, C. “Data Modeling for the Business - A Handbook for aligning the Business with IT using High-Level Data Models”. Editora Technics Publications, LLC, ISBN: 978-0977140077, 288 p., 2009

HOLLINK, V., TSIKRIKA, T., VRIES, A. P. “Semantic search log analysis: A method and a study on professional image search”. *Journal of the American Society for Information Science and Technology (JASIST)*, Vol. 62, Nr. 4, pp. 691-713, 2011.

INMON, WILLIAM H. “Como construir o Data Warehouse”. 2ª Edição, Editora Campus, São Paulo, Brasil, 266p., 1997.

ISLAM, A., INKPEN, D., KIRINGA, I. “Applications of corpus-based semantic similarity and word segmentation to database schema matching”. *The VLDB Journal*, Vol. 17, Nr. 5, pp. 1293-1320, 2008.

KASHYAP, V., BUSSLER, C., MORAN, M. “The Semantic Web - Semantics for Data and Services on the Web. Series: Data-Centric Systems and Applications”. Springer-Verlag, ISBN: 978-3-540-76451-9, 2008.

KATIFORI, A., HALATSIS, C., LEPOURAS, G. *et al.*: “Ontology visualization methods - a survey”. *ACM Computing Surveys (CSUR)*. Vol. 39, Nr. 4, Artigo 10, 2007.

LEE, J., SIAU, K., HONG, S. “Enterprise Integration with ERP and EAI”. *Communications of the ACM*. Vol. 46, Nr. 2, pp. 54-60, 2003.

LEI, Y., UREN, V. S., MOTTA, E. “Semsearch: A search engine for the semantic web*”.* In: S. Staab e V. Svátek,(eds), 5th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks, Podebrady, República Tcheca. Lect. Notes in Computer Science, Springer-Verlag, Vol. 4248. pp. 238-245, 2006.

LOPES, A. C. M., Rio-Torto, G., 2007. Semântica. Lisboa, Editorial Caminho.

MARCHIONINI, G. “Exploratory Search: From finding to understanding”. *Communications of the ACM*. Vol. 49, Nr.4, pp. 41-46, 2006.

MAEDCHE, A., STAAB, S.: “Comparing Ontologies— Similarity Measures and a Comparison Study”. Internal *Report No. 408, Institute AIFB*, University of Karlsruhe, Germany. Março 2001.

MAEDCHE, A., STAAB, S. “Measuring similarity between ontologies”. In: Proceedings *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pp. 251-263, 2002.

MALAMUD, C., *et. al*: Open Government Data Principles. Publicado em 8 de dezembro de 2007, Disponível on-line em http://www.opengovdata.org/home/8principles, Último acesso em 20/04/2011.

MANGOLD, C. A. “Survey and classification of semantic search approaches”. *International Journal of Metadata Semantics and Ontology*, Vol. 2, Nr. 1, 2007.

MELLO, R. S., HEUSER, C. A. “Binxs: A process for integration of xml schemata”. In: Pastor, O., Cunha, J. F. (eds.), *17th International Conference Advanced Information Systems Engineering*, Porto, Portugal, June 13-17, 2005. Lecture Notes in Computer Science, vol. 3520. Springer, pp. 151-166, 2005.

MELLO, R. S. *Uma abordagem Bottom-UP para a integração semântica de esquemas XML*. Tese de Doutorado. Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação, 2002.

MERRIAM-WEBSTER, INC, *Merriam-Webster's Collegiate Dictionary*, 11ª edição. Editora Merriam-Webster, Inc. 2005

MIHALCEA, R., CORLEY, C., STRAPPARAVA, C. “Corpus-based and knowledge-based measures of text semantic similarity”. In: A. Cohn (ed), *Proceedings of the 21st National Conference on Artificial Intelligence.* AAAI Press, Boston, Massachusetts, July 16 - 20, 2006. Vol. 1, pp. 775-780. 2006

MRABET, Y., BENNACER, N., PERNELLE, N., THIAM, M. “Supporting Semantic Search on Heterogeneous Semi-structured Documents*”*. In: *Proceedings of the 22nd International Conference on Advanced Information Systems Engineering*, Springer-Verlag Berlin, Heidelberg, LNCS, vol. 6051, pp. 224-229, 2010.

MYERS, M .D. “Qualitative Research in Information Systems”. MIS Quarterly. Vol. 21, Nr. 2, pp. 241-242, 1997. Disponível em www.qual.auckland.ac.nz. Último acesso em 10/11/2010.

NAVARRO, G. A “Guided Tour to Approximate String Matching”. *ACM Computing Surveys*, Vol. 33, No. 1, ACM New York, NY, USA, pp. 31-88, 1999.

NIGEL, S., HALL, W., BERNERS-LEE, T. “The Semantic Web Revisited”. *IEEE Intelligent Systems,* Vol. 21, Nr. 3, IEEE Educational Activities Department Piscataway, NJ, USA, pp.96-101, 2006

NATIONAL INFORMATION STANDARDS ORGANIZATION. “Understanding Metadata”. ISBN: 1-880124-62-9. *NISO Press*, USA, 2004. Disponível em http://www.niso.org/publications/press/UnderstandingMetadata.pdf. Último acesso em 01/02/2011.

NOY, N. F. “Semantic integration: a survey of ontology-based approaches”. *ACM SIGMOD Record*, Vol. 33, Nr. 4, ACM New York, NY, USA, pp. 65-70, 2004

NOY, N. F., MCGUINNESS, D. L. “*Ontology Development 101: A Guide to Creating Your First Ontology*”. Disponível em http://protege.stanford.edu/publications/ontology_development/ontology101.pdf. Último acesso em 01/05/2011.

PATRICK, P. “Impact of SOA on Enterprise Information Architectures”. In: *Proceeding of The 2005 ACM SIGMOD International Conference on Management of Data*, ACM New York, NY, USA, pp.844-849, 2005

PETRAKIS, E. “Information retrieval by semantic similarity”. *International Journal on Semantic Web and Information Systems*. Vol. 2 Nr. 3, pp.55-73, 2006

PRICEWATERHOUSECOOPERS. “Spinning a data Web”. *Technology Forecast - A quarterly journal*. Spring, 2009

PY, H., CASTRO, L., BAIÃO, F. A., TANAKA, A. “A service-based approach for data integration based on business process models”. In: J. Cordeiro e J. Filipe, (eds), ICEIS 2009 - *Proceedings of the 11th International Conference on Enterprise Information Systems*, Volume DISI, Milan, Italy, May 6-10, pp. 222-227, 2009

RAHM, E., BERNSTEIN, P. A. “A survey of approaches to automatic schema matching”. *The VLDB Journal — The International Journal on Very Large Data Bases*, Vol. 10 Nr. 4, pp. 334–350, 2001

REEVE, L., HAN, H. “Survey of Semantic Annotation Platforms”. In: *Proceedings of the 2005 ACM Symposium on Applied Computing*, ACM Press, pp 1634-1638, 2005

SATTLER, K. “Data Quality Dimensions”. In: Liu, L., Ozsu, M. T. (eds), *Encyclopedia of Database Systems*, Springer, pp. 612-615, 2008.

SHETH, A. P. “Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics”. *Interoperating Geographic Information Systems*, pp. 5-30, 1998

SICILIA, M. “Metadata, semantics, and ontology: providing meaning to information resources”. *Internationa. Journal Metadata, Semantics and Ontologies*, Vol. 1, No. 1, 2006

SIMMHAN, Y. L. *Provenance Framework in Support of Data Quality Estimation*. Tese de Doutorado, Indiana University, 2007.

SOARES, P., TANAKA, A., BAIÃO, F. A. "Estudo dos Principais Conceitos sobre Integração de Dados Geoespaciais", *RelaTE-DIA: Relatórios Técnicos do Departamento de Informática Aplicada (DIA) – Universidade Federal do Estado do Rio de Janeiro (UNIRIO)*, No. 00018/2010, Dezembro, 2010.

SLTI-MP. Apresentação "*Acompanhamento do Planejamento da INDA*" realizada em Brasília, 19 de Abril de 2011 pela Secretaria de Logística e Tecnologia da Informação (SLTI). Disponível para download em http://wiki.gtinda.ibge.gov.br/GetFile.aspx?Page=Agenda-do-GT&File=15%20-%20Encontro%20de%20acompanhamento%20da%20INDA%20-%2019.04.2011.odp. Último acesso em 17/05/2011.

SMITH, K., MORSE, M., MORK, P., *et al.* “The role of schema matching in large enterprises”. CIDR 2009, *Fourth Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, January 4-7, 2009

TAYI, G. K., BALLOU, D. P. “Examining Data Quality*”. Communications of the ACM,* Vol. 41 Nr. 2, pp. 54–57, 1998.

THIAGARAJAN, R., MANJUNATH, G., STUMPTNER, M. *Computing Semantic Similarity Using Ontologies*. In: HP Labs Technical Report HPL-2008-87, Jul. 2008

UREN, V. S., LEI, Y., MOTTA, E. “Semsearch: Refining semantic search”. In: *ESWC, S. Bechhofer, M. Hauswirth, J. Ho_mann, and M. Koubarakis, Eds. Lecture Notes in Computer Science,* vol. 5021. Springer, pp. 874-878, 2008.

USCHOLD, M., GRUNINGER, M. “Ontologies and semantics for seamless connectivity”. *ACM SIGMOD Record*, Vol. 33, Nr. 4, pp. 58-64, 2004

VCGE. Vocabulário Controlado de Governo Eletrônico. Versão 1, Novembro 2010.

WACHE, H., VOEGELE, T., VISSER, U., *et al.*: “Ontology-Based Integration of Information - A Survey of Existing Approaches*”.* In *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, Seattle, WA, August 4-5, pp 108-118, 2001.

WAINER, J.: “Métodos de pesquisa quantitativa e qualitativa para a ciência computação”. Tomasz Kowaltowski, Karin Breitman. (eds). *Atualização em informática 2007*. Sociedade Brasileira de Computação e Editora PUC-Rio, pp. 221-262, 2007.

WEISZFLOG, W. Michaelis Moderno Dicionário Da Língua Portuguesa. Editora Melhoramentos. 2006.

WRIGLEY, S. N., REINHARD, D., ELBEDWEIHY, K., *et al.* “Methodology and Campaign Design for the Evaluation of Semantic Search Tools”. In: *Proceedings of the 3rd International Semantic Search Workshop*, ACM New York, NY, USA, 2010

W3C Brasil. *Dados Abertos Governamentais.* Disponível em http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf. Último acesso em 07/fevereiro/2011.

W3C Brasil. *Melhorando o acesso ao governo com o melhor uso da web*. Disponível em http://www.w3c.br/divulgacao/pdf/gov-web.pdf. Último acesso em 11/março/2011.

YIN, R. K. *Estudo de Caso: Planejamento e Métodos.* 3 edição, Porto Alegre, Ed. Bookman, 2005.

YUAN A., BORGIDA, A., MYLOPOULOS, J. “Discovering the Semantics of Relational Tables through Mappings”, *Journal on Data Semantics VII*, LNCS 4244, pp. 1-32, 2006.

ZIEGLER, P., DITTRICH, K. “Data Integration - Problems, Approaches and Perspectives”. In: Krogstie, J., Opdahl, A. L., Brinkkemper S. (eds), *Conceptual Modelling in Information Systems Engineering*, p. 39–58. Springer, 2007.

Anexo I – Configuração FTS

Experimento

```
CREATE TEXT SEARCH DICTIONARY public.pt_ispell (
   TEMPLATE = ispell,
   dictfile = 'portuguese', afffile = 'portuguese', stopwords = 'portuguese');

CREATE TEXT SEARCH CONFIGURATION public.fts_meta (  PARSER = "default");

ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR asciihword WITH pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR asciiword WITH pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR email WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR file WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR float WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR host WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword WITH pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword_asciipart WITH pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword_numpart WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword_part WITH pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR int WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR numhword WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR numword WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR sfloat WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR uint WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR url WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR url_path WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR version WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR word WITH pt_ispell,portuguese_stem;
```

Estudo de Caso

```
CREATE TEXT SEARCH DICTIONARY public.pt_hunspell (
   TEMPLATE = ispell,
   dictfile = 'pt_hunspell', afffile = 'pt_hunspell', stopwords = 'portuguese');

CREATE TEXT SEARCH DICTIONARY public.pt_ispell (
   TEMPLATE = ispell,
   dictfile = 'portuguese', afffile = 'portuguese', stopwords = 'portuguese');

CREATE TEXT SEARCH CONFIGURATION public.fts_meta ( PARSER = "default");

ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR asciihword
WITH pt_hunspell,pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR asciiword WITH
pt_hunspell,pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR email WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR file WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR float WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR host WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword WITH
pt_hunspell,pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword_asciipart
WITH pt_hunspell,pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword_numpart
WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR hword_part
WITH pt_hunspell,pt_ispell,portuguese_stem;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR int WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR numhword WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR numword WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR sfloat WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR uint WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR url WITH simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR url_path WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR version WITH
simple;
ALTER TEXT SEARCH CONFIGURATION public.fts_meta ADD MAPPING FOR word WITH
pt_hunspell,pt_ispell,portuguese_stem;
```

Anexo II – Modelos de Dados do Experimento

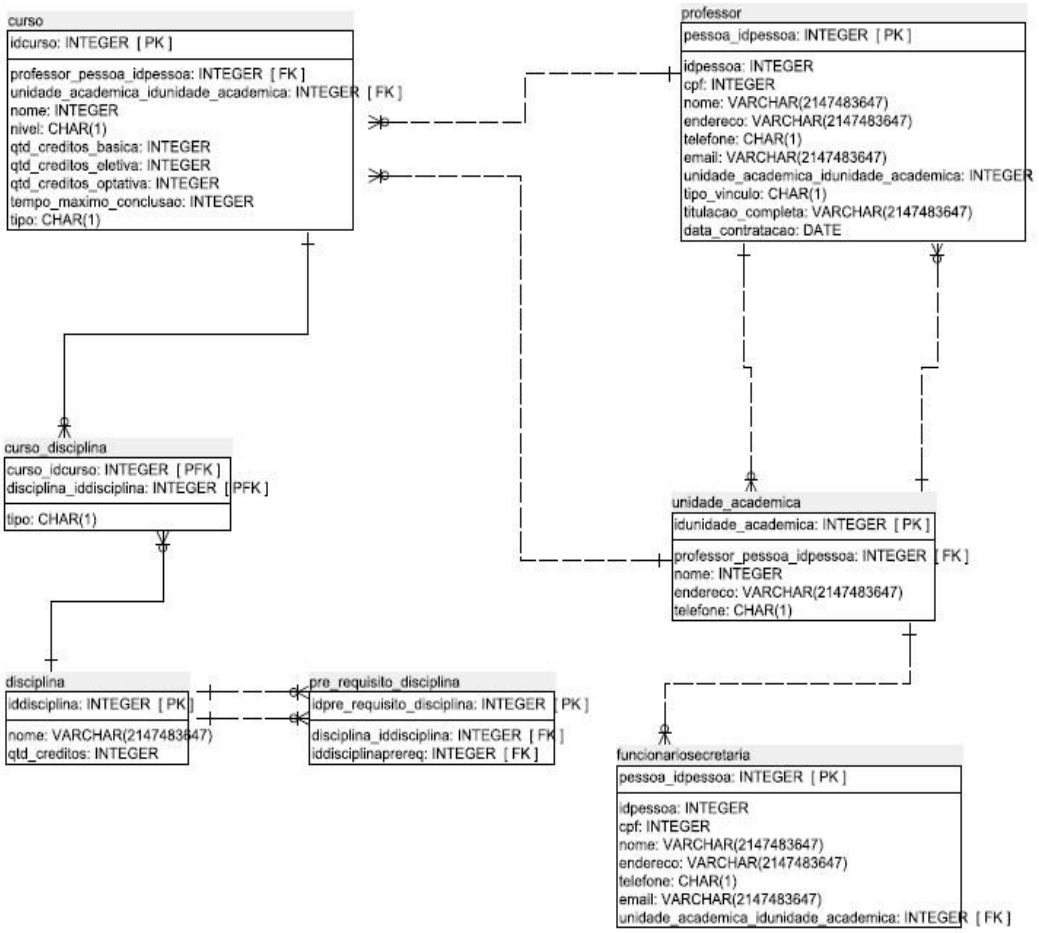


Figura A2.1 – Modelo de dados do esquema gcad1

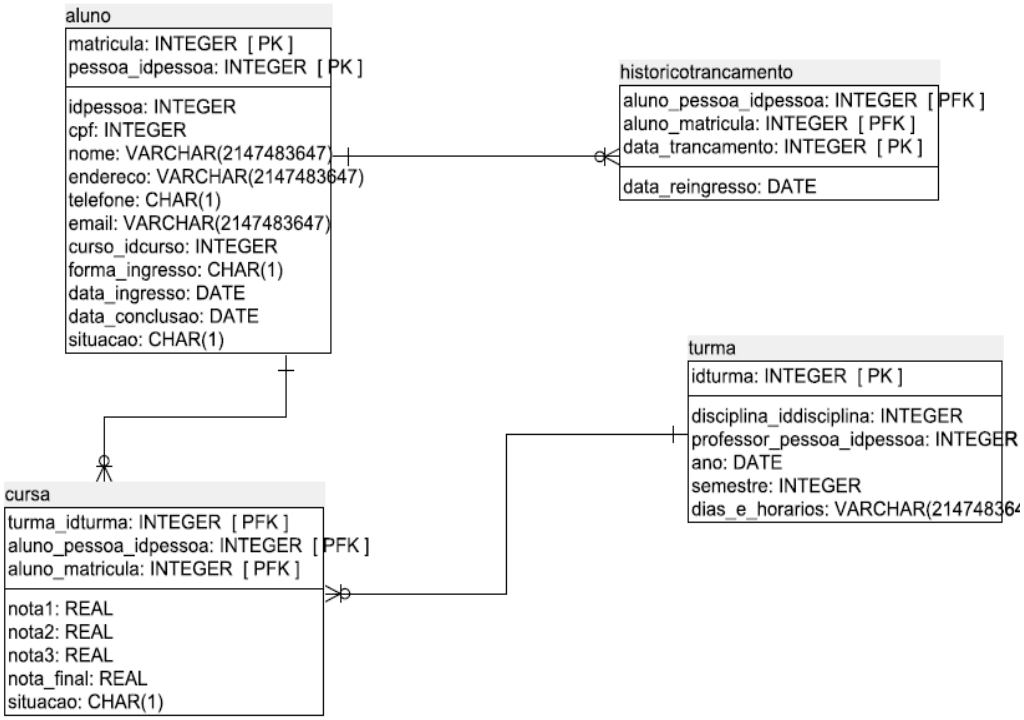


Figura A2.2 – Modelo de dados do esquema gcad2

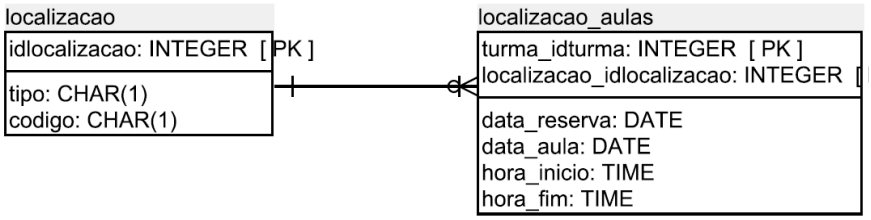


Figura A2.3 – Modelo de dados do esquema gcad2

```
xml version="1.0" encoding="ISO-8859-1" ?>
<xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema
">

<!-- definition of simple elements -->
<xs:element name="nome" type="xs:string"/>
<xs:element name="curso"
type="xs:string"/>
<xs:element name="creditos">
  <xs:simpleType>
    <xs:restriction
base="xs:positiveInteger">
      <xs:minInclusive value="5"/>
      <xs:maxInclusive value="10"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="ementa"
type="xs:string"/>
<xs:element name="codigo"
type="xs:string"/>
<xs:element name="professor"
type="xs:string"/>
<xs:element name="diaSemana"
type="xs:string"/>
<xs:element name="horaInicio"
type="xs:string"/>
<xs:element name="horaFim"
type="xs:string"/>
<xs:element name="modalidade">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:pattern value="presencial|a
distancia"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="tipo">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:pattern
value="basica|eletiva|optativa"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="titulo"
type="xs:string"/>
<xs:element name="autor"
type="xs:string"/>
<xs:element name="veiculo"
type="xs:string"/>
<xs:element name="edicao"
type="xs:positiveInteger"/>
<xs:element name="licenca">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:pattern
value="proprietario|livre"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="ambiente">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:pattern value="windows|linux"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>

<!-- definition of attributes -->
<xs:attribute name="tipo"
type="xs:string"/>
```

```
<!-- definition of complex elements -->
<xs:element name="gradeHorario">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="diaSemana"/>
      <xs:element ref="horaInicio"/>
      <xs:element ref="horaFim"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="bibliografia">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="titulo"/>
      <xs:element ref="autor" />
      <xs:element ref="veiculo" />
      <xs:element ref="edicao" />
    </xs:sequence>
    <xs:attribute ref="tipo" use="required"/>
  </xs:complexType>
</xs:element>

<xs:element name="programa">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="nome" />
      <xs:element ref="versao" />
      <xs:element ref="ambiente" />
      <xs:element ref="licenca" />
    </xs:sequence>
  </xs:complexType>
</xs:element>

<xs:element name="disciplina">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="nome" maxOccurs="1"/>
      <xs:element ref="curso" />
      <xs:element ref="creditos" maxOccurs="1"/>
      <xs:element ref="ementa" maxOccurs="1"/>
      <xs:element ref="codigo" maxOccurs="1"/>
      <xs:element ref="professor" />
      <xs:element ref="modalidade"
maxOccurs="1"/>
      <xs:element ref="tipo" maxOccurs="1"/>
      <xs:element ref="gradeHorario" />
      <xs:element ref="bibliografia"
maxOccurs="unbounded"/>
      <xs:element ref="programa"
maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

</xs:schema>
```

Figura A2.4 – Esquema XML para descrição de disciplinas no site

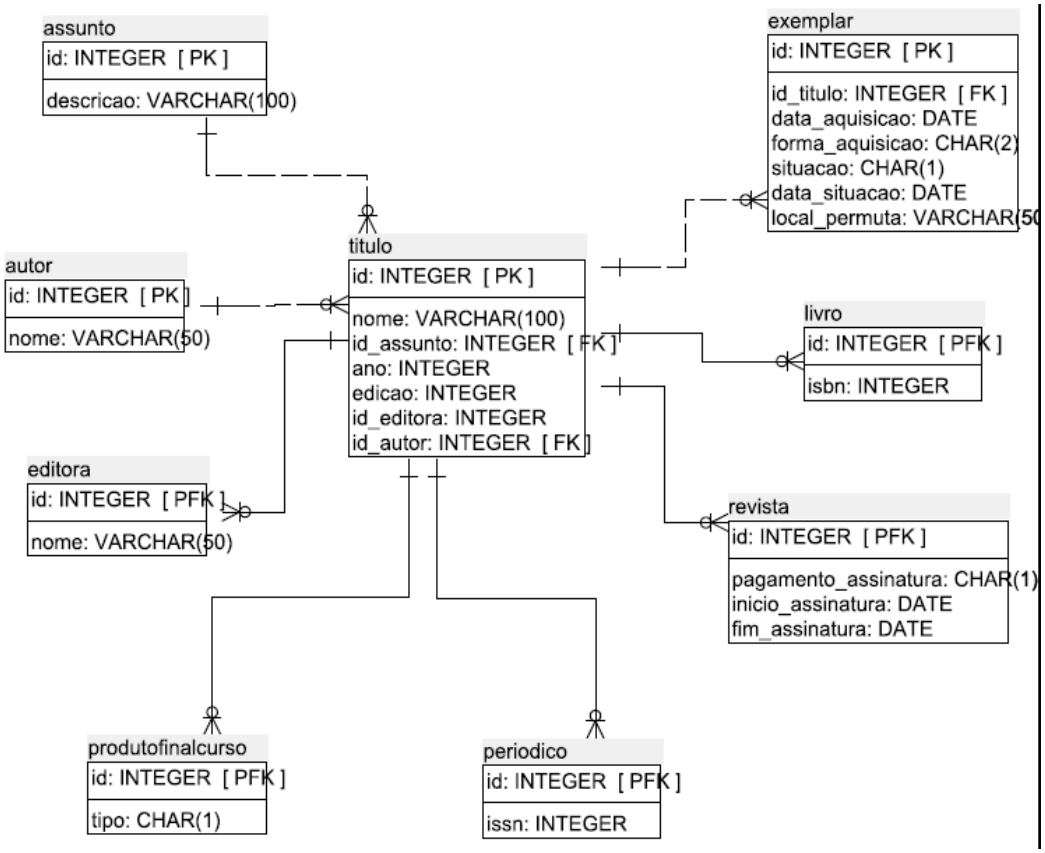


Figura A2.5 – Modelo de dados do esquema bib_titulos

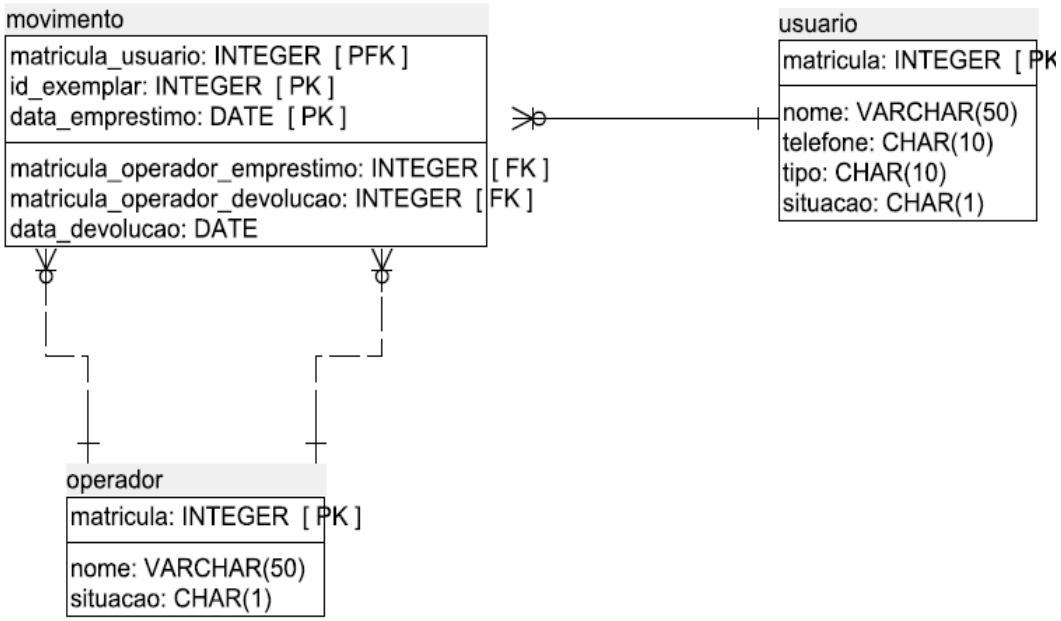


Figura A2.6 – Modelo de dados do esquema bib_movimento

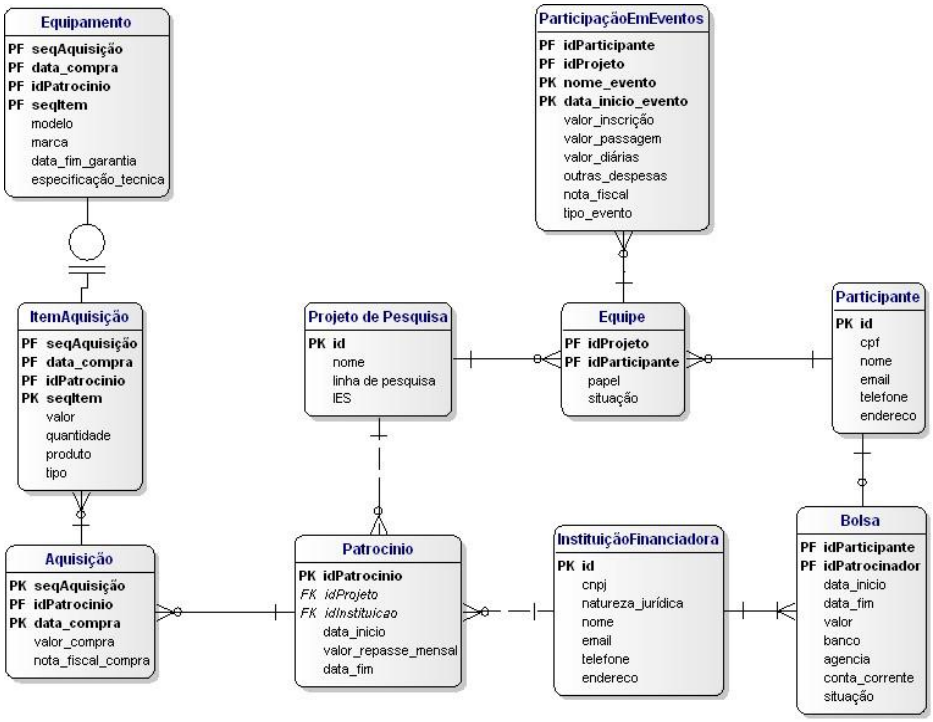


Figura A2.7 – Modelo de dados do FinPesq

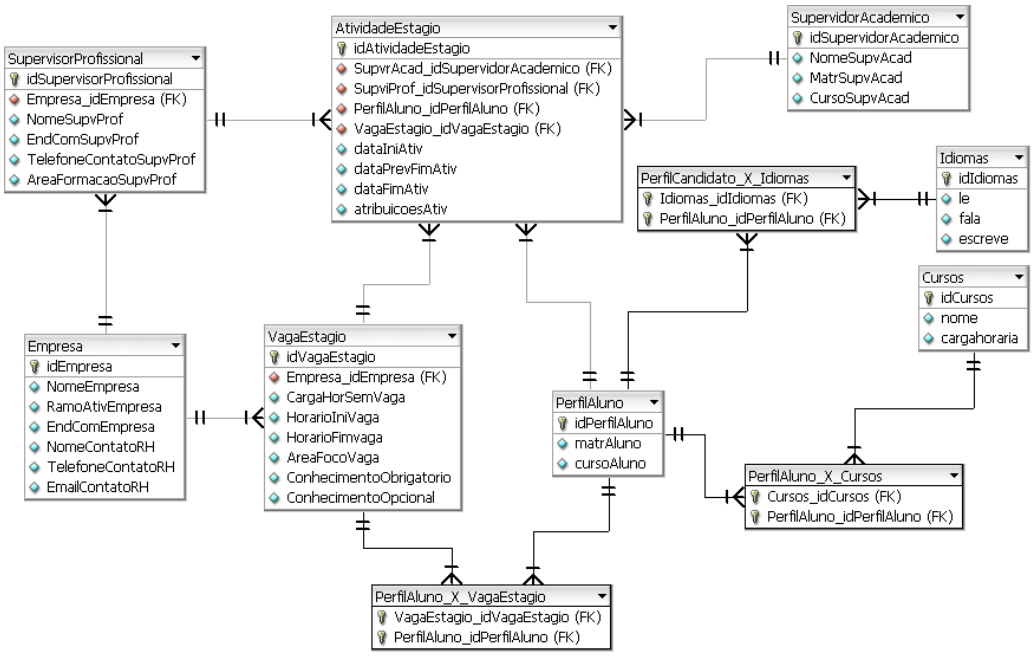
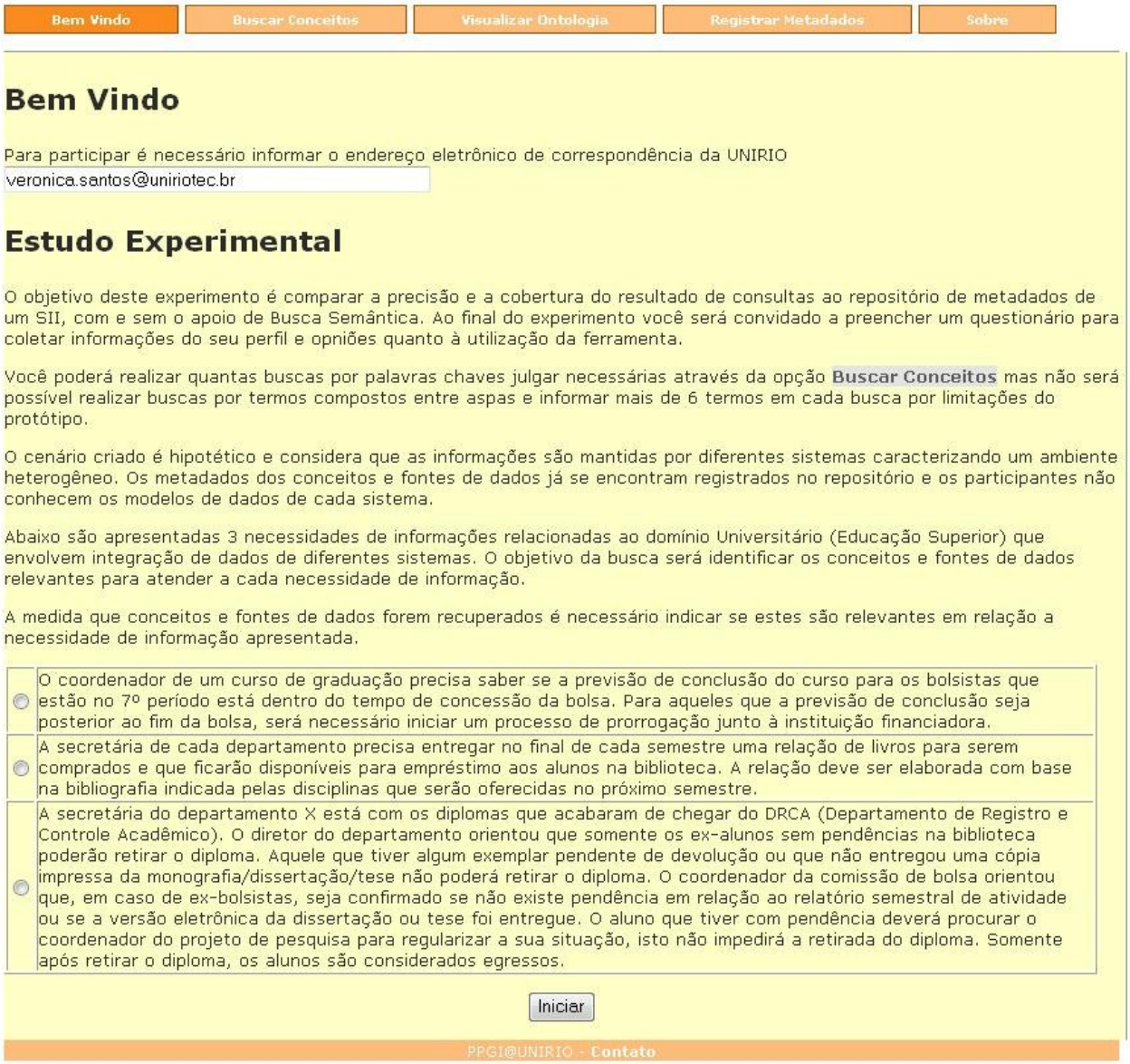


Figura A2.8 – Modelo de dados do esquema gcad2

Na tela **Bem Vindo** é necessário informar o endereço de correspondência, selecionar a necessidade de informação a ser tratada e clicar no botão **Iniciar**.

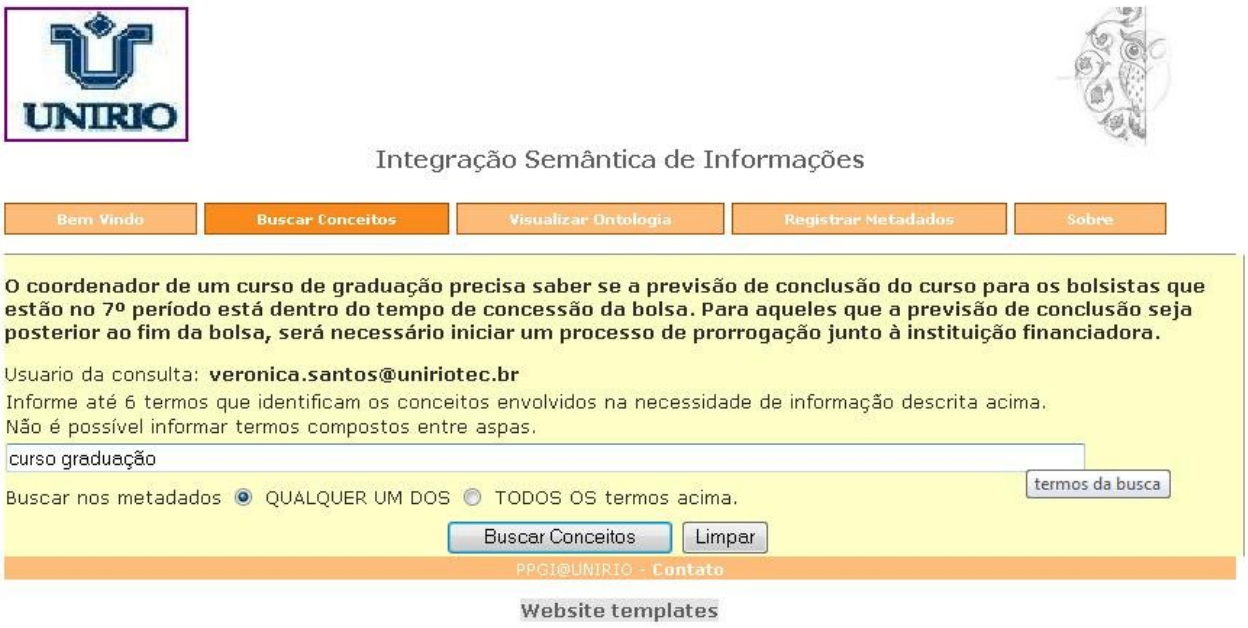


Cada participante pode realizar quantas buscas por palavras chaves julgar necessárias através da opção **Buscar Conceitos**. Em cada consulta é necessário informar os termos que expressam conceitos envolvidos na necessidade de informação apresentada.

Por limitações do protótipo não será possível realizar busca por termos compostos entre aspas e informar mais de 6 termos em cada consulta.

É possível escolher entre buscar TODOS (AND) ou QUALQUER UM (OR) os termos no repositório de metadados.

Clique no botão **Buscar Conceitos**.



Uma camada de reformulação da consulta recebe os termos da interface, realiza a busca na ontologia de domínio por recursos (conceitos, relações e atributos) que possuam rótulos coincidentes com estes termos.

Para cada recurso identificado é recuperado o seu URI e os respectivos rótulos. Além disto são recuperados o URI e os respectivos rótulos de outros recursos seguindo os seguintes critérios:
1) Se o recurso encontrado for um conceito então são acrescentados nos resultados:
- seus subconceitos, se houverem;
- seus superconceitos com os respectivos conceitos irmãos e atributos herdados;
- os relacionamentos que o conceito participa e os respectivos conceitos relacionados;
- os atributos.
2) Se o recurso encontrado for um conceito que possui algum conceito equivalente então são acrescentados nos resultados para cada conceito equivalente:
- seus subconceitos, se houverem;
- seus superconceitos com os respectivos conceitos irmãos e atributos herdados;
- os relacionamentos que o conceito equivalente participa e os respectivos conceitos relacionados;
- os atributos.
3) Se o recurso encontrado for um conceito que é formado a partir da união de outros conceitos então são acrescentados nos resultados cada conceito que é parte desta união.
4) Se o recurso encontrado for um conceito que é participa de uma união com outros conceitos para formar um novo conceito então são acrescentados nos resultados cada conceito que é parte desta união e o conceito que representa o resultado da união.
5) Se o recurso encontrado for um relacionamento então são acrescentados nos resultados os conceitos que participam do relacionamento.
6) Se o recurso encontrado for um relacionamento que possui algum relacionamento equivalente então são acrescentados nos resultados os conceitos que participam de cada relacionamento equivalente.
7) Se o recurso encontrado for um atributo então são acrescentados nos resultados os conceitos que possuem este atributo.

Se o conjunto de recursos recuperados para um termo informado possui algum recurso em comum com o conjunto de recursos de outros termos, estes são considerados como termos diretamente ligados. Se o conjunto de recursos associados a um termo é um subconjunto dos recursos de outro termo, então o primeiro termo está contido no segundo. Nestes dois casos os recursos compartilhados são indicados em negrito na tela.

Se o conjunto de recursos associados a um termo é exatamente igual ao de outro termo, estes são considerados como equivalentes e os recursos são relacionados a somente um dos termos na tela.

Os rótulos dos recursos recuperados da ontologia são apresentados ao usuário, permitindo a seleção da expansão desejada a ser aplicada. Os rótulos selecionados são acrescentados no conjunto de palavras chaves da busca e os respectivos ponteiros para o recurso usados na busca por metadados anotados.

Clique no botão **Buscar Metadados**.

O coordenador de um curso de graduação precisa saber se a previsão de conclusão do curso para os bolsistas que estão no 7º período está dentro do tempo de concessão da bolsa. Para aqueles que a previsão de conclusão seja posterior ao fim da bolsa, será necessário iniciar um processo de prorrogação junto à instituição financiadora.

Usuario da consulta: **veronica.santos@uniriotec.br**
Resultados retornados pela busca dos termos **curso graduação** na ontologia.

curso **está ligado a** graduação

Selecione o(s) tipo(s) de expansão e o(s) termo(s) que serão acrescentados para realizar a busca no repositório de metadados.

Termo	Rótulos	Tipo de Expansão
curso	☑ **Curso|**	**Conceito**
	☑ **.Doutorado|**	**.Subconceito**
	☑ **.Extensão|**	**.Subconceito**
	☑ **.Graduação|**	**.Subconceito**
	☑ **.MBA|**	**.Subconceito**
	☑ **.Mestrado|**	**.Subconceito**
	☑ **.Pós Graduação Lato Sensu|**	**.Subconceito**
	☐ .Coordena|	.Relacionamento
	☐ ..Coordenador|Diretor|Gestor|	..Conceito Relacionado
	☐ .é Coordenado por|	.Relacionamento
	☐ ..Coordenador|Diretor|Gestor|	..Conceito Relacionado
	☐ .Pertecence a Grade do|faz parte da grade do|	.Relacionamento
	☐ ..Disciplina|	..Conceito Relacionado
	☐ .Pertence a|	.Relacionamento
	☐ ..Unidade Acadêmica|	..Conceito Relacionado
	☐ .Possui|	.Relacionamento
	☐ ..Unidade Acadêmica|	..Conceito Relacionado
	☑ **.nome curso|**	**.Atributo**
	☑ **nome curso|**	**Atributo**
	☑ **.Curso|**	**.Conceito**
graduação	☑ **Graduação|**	**Conceito**
	☑ **.Curso|**	**.Superconceito**
	☑ **..Doutorado|**	**..Conceito Irmão**
	☑ **..Extensão|**	**..Conceito Irmão**
	☑ **..MBA|**	**..Conceito Irmão**
	☑ **..Mestrado|**	**..Conceito Irmão**
	☑ **..Pós Graduação Lato Sensu|**	**..Conceito Irmão**
	☑ **..nome curso|**	**..Atributo**
	☑ **Pós Graduação Lato Sensu|**	**Conceito**
	☑ **.Curso|**	**.Superconceito**
	☑ **..Doutorado|**	**..Conceito Irmão**
	☑ **..Extensão|**	**..Conceito Irmão**
	☑ **..Graduação|**	**..Conceito Irmão**
	☑ **..MBA|**	**..Conceito Irmão**
	☑ **..Mestrado|**	**..Conceito Irmão**
	☑ **..nome curso|**	**..Atributo**
	☑ **MBA|**	**Conceito Equivalente**
	☑ **.Curso|**	**.Superconceito**
	☑ **..Doutorado|**	**..Conceito Irmão**
	☑ **..Extensão|**	**..Conceito Irmão**
	☑ **..Graduação|**	**..Conceito Irmão**
	☑ **..Mestrado|**	**..Conceito Irmão**
	☑ **..Pós Graduação Lato Sensu|**	**..Conceito Irmão**
	☑ **..nome curso|**	**..Atributo**

Buscar Metadados | Limpar

A camada de reformulação aplica as modificações na consulta original, agregando os termos da ontologia e ponteiros (URI) dos recursos, interage com o repositório de metadados para recuperar os conceitos e fontes de informações associados através da similaridade léxica com os termos da consulta e existência de anotação semântica nos metadados associadas aos recursos selecionados.

Ao final o resultado é ordenado por relevância, considerando a existência de anotação semântica associada ao recurso, o critério de qualidade das fontes de informações e a similaridade léxica do conteúdo da descrição dos metadados em relação aos termos da consulta, para ser apresentado ao usuário.

Através da tela de apresentação dos resultados é possível visualizar o recurso associado a anotação semântica, se existir, clicando na opção **Visualizar anotação semântica**. Também é possível visualizar a estrutura do esquema clicando em **Visualizar esquema**.

A medida que conceitos e fontes de informações forem recuperados e apresentados é necessário indicar se estes são relevantes em relação a necessidade de informação que está sendo tratada. Na tela de resultado, o participante deve selecionar o(s) par(es) conceito / fontes de informações.

Caso o participante deseje realizar novas consultas para recuperar outros conceitos / fontes de informações que são relevantes então o botão **Realizar nova busca** deve ser clicado e o sistema retornará a tela inicial de busca (Buscar Conceitos).

Caso o participante tenha concluído a seleção de conceitos / fontes de informações relevantes então o botão **Concluir seleção** deve ser clicado para que o sistema direcione o usuário para a tela de seleção de outra necessidade de informação (Bem Vindo).

Resultados retornados pela busca dos termos **curso graduação** e recursos associados nos metadados.

Selecione o(s) conceitos(s) e fonte(s) de dados que atendem a necessidade de informação.

Selecionar	Conceitos e Fontes de Dados
1	Conceito: Cursos Descrição: cursos vinculados ao departamento. Podem ser de graduação, pós ou de extensão. Serviço: srv_curso Descrição: dados do curso (1) Esquema: gcad1 Descrição: professor e funcionario, departamento, cursos e suas disciplinas Fonte: gcad01 Tipo: PostgreSQL 8.4 Serviço: srv_gcad1 Descrição: acesso ao esquema gcad1 do sistema Gerencia Acadêmica Critério de Qualidade - Prioridade da Fonte: **1** Similaridade léxica dos termos da consulta: **0.209802** Visualizar anotação semântica | Visualizar esquema
2	Conceito: Cursos Descrição: cursos vinculados ao departamento. Podem ser de graduação, pós ou de extensão. Serviço: srv_curso Descrição: dados do curso (2) Esquema: public Descrição: esquema da descrição de disciplinas de cada curso com ementa e bibliografia. É publicado no site de cada departamento. Fonte: site departamento Tipo: XML Serviço: srv_site Descrição: obtem documentos xml do site Critério de Qualidade - Prioridade da Fonte: **2** Similaridade léxica dos termos da consulta: **0.210126** Visualizar anotação semântica | Visualizar esquema
3	Conceito: aluno de ***mestrado*** Descrição: Aluno que está regularmente inscrito em um ***curso*** de ***mestrado*** Serviço: Não há Descrição: Não há (1) Esquema: gcad2 Descrição: alunos, matrículas, inscrições em disciplinas e suas notas e turmas Fonte: gcad01 Tipo: PostgreSQL 8.4 Serviço: srv_gcad2 Descrição: acesso ao esquema gcad2 do sistema Gerencia Acadêmica Critério de Qualidade - Prioridade da Fonte: **1** Similaridade léxica dos termos da consulta: **0.131931** Visualizar anotação semântica | Visualizar esquema
12	Conceito: turma Descrição: uma turma contém pelo menos três estudantes e se refere as aulas de uma disciplina ministrada por um professor dentro de um horário. Está associada em um semestre/ano e as aulas são ministradas em salas, laboratórios ou auditórios Serviço: srv_turma Descrição: info sobre turmas em relação aos estudantes que fazem parte da mesma, professor que leciona e qual disciplina, horário, semestre, ano, etc ... (1) Esquema: gcad1 Descrição: professor e funcionario, departamento, ***cursos*** e suas disciplinas Fonte: gcad01 Tipo: PostgreSQL 8.4 Serviço: srv_gcad1 Descrição: acesso ao esquema gcad1 do sistema Gerencia Acadêmica Critério de Qualidade - Prioridade da Fonte: **1** Similaridade léxica dos termos da consulta: **0.0046546** Não há anotação semântica | Visualizar esquema
13	Conceito: Secretaria Descrição: A secretaria é uma unidade administrativa que apóia a gestão da unidade acadêmica Serviço: srv_secretaria Descrição: lista de funcionários administrativos (1) Esquema: gcad1 Descrição: professor e funcionario, departamento, ***cursos*** e suas disciplinas Fonte: gcad01 Tipo: PostgreSQL 8.4 Serviço: srv_gcad1 Descrição: acesso ao esquema gcad1 do sistema Gerencia Acadêmica Critério de Qualidade - Prioridade da Fonte: **1** Similaridade léxica dos termos da consulta: **0.0046546** Visualizar anotação semântica | Visualizar esquema
14	Conceito: Disciplinas Descrição: as disciplinas são oferecidas pelo departamento mas podem ser ***cursadas*** por alunos de outros departamentos e também podem ser oferecidas para compor a grade de ***cursos*** de outros departamentos. Serviço: srv_disciplinas Descrição: todas as info sobre disciplinas (2) Esquema: public Descrição: esquema da descrição de disciplinas de cada ***curso*** com ementa e bibliografia. É publicado no site de cada departamento. Fonte: site departamento Tipo: XML Serviço: srv_site Descrição: obtem documentos xml do site Critério de Qualidade - Prioridade da Fonte: **2** Similaridade léxica dos termos da consulta: **0.0874792** Não há anotação semântica | Visualizar esquema

Realizar nova busca | Concluir seleção | Limpar