Um Servidor de Ontologias para apoio a Sistemas de Biodiversidade

Jaudete Daltio[1], **Claudia M. Bauzer Medeiros**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Caixa Postal 6176 – 13084-971 – Campinas – SP – Brasil

{ra049240,cmbm}@ic.unicamp.br

Nível: Mestrado
Ano de ingresso no programa: Agosto/2005
Época esperada de conclusão: Agosto/2007
Etapas já concluídas: Exame de qualificação: 03/05/2006

Abstract. *Biodiversity research requires associating data about living beings and their habitats, integrating from geographical features to domain specifications, often through ontologies. In this context are the so-called Biodiversity Information Systems, new management solutions that allow researchers to analyze species characteristics and their interactions. The goal of this project is to specify and develop an ontology web service that can be used for different biodiversity systems. The main contributions of this work are: specification of the requirements of an ontology service; and the specification and the implementation of an ontology server.*

Resumo. *A pesquisa em biodiversidade requer correlacionar dados sobre seres vivos e seus habitats, integrando desde relacionamentos espaciais a especificações de domínio, frequentemente usando ontologias. Nesse contexto estão os Sistemas de Biodiversidade, novas soluções de gerenciamento que permitem aos pesquisadores analisar as características das espécies e suas interações. O objetivo deste projeto é especificar e desenvolver um serviço web para ontologias, que possa ser usado por diferentes sistemas de biodiversidade. Dentre as contribuições deste trabalho estão a especificação das necessidades de um serviço de ontologias; e a especificação e implementação deste serviço.*

1. Introdução e Motivação

A pesquisa em biodiversidade é um assunto de grande destaque. Dada a grande diversidade de espécies existentes, a coleta, o processamento e a descrição de dados de biodiversidade por meios tradicionais tornam-se cada vez mais inviáveis. Dentre os maiores desafios enfrentados estão a identificação e avaliação de descontinuidades críticas no domínio de biodiversidade (tanto taxonômicas quanto geográficas) e a mineração de dados em coleções existentes.

Esse cenário tem motivado diversos esforços na coleta de dados, resultando em uma grande quantidade de informações que requer novas soluções de gerenciamento que permitam aos pesquisadores analisar as características das espécies e suas interações.

Com esse objetivo estão em desenvolvimento os *Sistemas de Informação de Biodiversidade* [Torres 2004]. Um grande desafio desses sistemas está no gerenciamento de coleções de dados, cujas origens são freqüentemente desestruturadas e heterogêneas.

Nesse contexto está o *WeBios* [web 2005], um projeto em desenvolvimento conduzido por pesquisadores dos Institutos de Computação e de Biologia da UNICAMP. Esse projeto tem o objetivo de prover um sistema que suporte consultas exploratórias multimodais sobre fontes de dados heterogêneas de biodiversidade (dados textuais de espécies, imagens, dados geográficos, ontologias e anotações), acessadas via serviços *Web*.

Tais serviços poderiam resolver problemas relacionados à heterogeneidade e variação de qualidade intrínsecas às fontes de dados usadas. Além de variações na coleta dos dados, fatores externos como a classificação de espécies e modelos ecológicos variam no tempo, refletindo a evolução do conhecimento científico no mundo real. Imagens de má qualidade ou dados escassos/incertos são outros problemas encontrados.

Uma das soluções apontadas para resolver alguns desses problemas é o uso de ontologias como forma de permitir o entendimento de termos e definições, e seus relacionamentos. No entanto, grupos distintos de pesquisadores usam ontologias diferentes, e há variações de ontologias para cada domínio do conhecimento. O objetivo desta dissertação é especificar e implementar um servidor de ontologias, capaz de realizar a integração de ontologias que compartilhem domínios similares ou complementares, provendo um resultado mais expressivo às consultas formuladas.

A seção 2 apresenta alguns conceitos teóricos importantes para a contextualização do projeto. A seção 3 descreve o projeto proposto nesta dissertação. A seção 4 apresenta algumas iniciativas similares e uma análise comparativa de suas características mais relevantes. A seção 5 descreve os resultados esperados e as contribuições previstas.

2. Fundamentação Teórica

A análise integrada de dados heterogêneos acarreta uma grande variedade de problemas. Ao lado das técnicas de integração de sistemas existe a integração semântica, cuja abordagem de maior destaque é a integração de ontologias. Ontologias são expressões ou descrições de um modelo abstrato de termos, relacionados entre si. Elas modelam uma parte da realidade, suas relações taxonômicas e não-taxonômicas e as restrições aplicadas a essas entidades de domínio, com a finalidade de definir um entendimento comum. As ontologias têm se tornado uma parte importante em aplicações industriais e acadêmicas.

A integração de ontologias consiste em reconciliar diferenças entre ontologias para alcançar interoperabilidade entre aplicações. Isso inclui a descoberta e a especificação de relacionamentos entre as entidades que compõem as ontologias. Dentre as abordagens para essa integração, citamos:
Mapeamento de ontologias [Kalfoglou and Schorlemmer 2003]: refere-se à identificação de conceitos ou relacionamentos idênticos entre diferentes ontologias, especificando como se relacionam. As ontologias originais não se modificam, sendo inseridos axiomas adicionais para descrever o relacionamento entre os conceitos mapeados.
Alinhamento de ontologias [Ehrig et al. 2004]: pode ser visto como um conjunto de mapeamentos de elementos e requer entendimento dos significados dos conceitos. As ontologias são mantidas separadas, mas pelo menos uma das originais é adaptada de

forma que a conceitualização e o vocabulário se emparelhem, sobrepondo partes das ontologias. O resultado desse processo é trazer as ontologias para um "acordo mútuo", embora possam descrever diferentes partes do domínio em distintos níveis de detalhe.
União (merging) de ontologias [Ehrig et al. 2004]: ocorre quando o processo de combinação, com base em duas ou mais ontologias já existentes, resulta na construção de uma nova ontologia.

Abordagens para integração de ontologias são, em geral, baseadas na análise de todos os dados e no reconhecimento de partes que "casam" umas com as outras. Esse casamento pode se basear na identificação de partes idênticas ou na identificação de elementos de relacionamentos (por exemplo, parte-de). Segundo [Shvaiko and Euzenat 2004], existem duas principais classificações dessas técnicas:

Técnicas de combinação em nível de elemento [Rahm and Bernstein 2001, Shvaiko and Euzenat 2004]: computam o mapeamento dos elementos analisando as entidades isoladamente, ignorando seus relacionamentos. Nessa categoria estão: (1) técnicas baseadas em *string*: baseiam-se na similaridade de nome e descrição dos elementos da ontologia; quanto mais similares as *strings*, mais provável que denotem os mesmos conceitos; (2) técnicas baseadas em linguagem: consideram *strings* como palavras em alguma linguagem natural e exploram as propriedades morfológicas dos termos, como a detecção da forma básica das palavras e a eliminação de artigos, preposições e conjunções; (3) técnicas baseadas em restrições: analisam as restrições aplicadas às definições das entidades, como seus tipos de dados, cardinalidades e instâncias (duas classes com as mesmas instâncias possuem grande chance de denotarem conceitos similares); (4) recursos lingüísticos: usam dicionários de domínio na tentativa de combinar palavras baseando-se nos relacionamentos lingüísticos entre elas (sinônimos, hiponônimos).

Técnicas em nível de estrutura [Abels et al. 2005, Shvaiko and Euzenat 2004, Rahm and Bernstein 2001]: computam o mapeamento dos elementos analisando como as entidades aparecem juntas na estrutura. Nessa categoria estão: (1) técnicas baseadas em grafos: consideram uma ontologia como uma estrutura de grafo e buscam estruturas similares em duas ontologias; (2) técnicas baseadas em taxonomia: algoritmos de grafos que consideram apenas as relações de especialização, identificando classes idênticas a partir de seus atributos e das classes com as quais elas se relacionam.

3. Trabalho Proposto na Dissertação

O projeto *Webios*, dentro do qual está a dissertação, permite ao usuário consultar de forma integrada fontes heterogêneas de dados de biodiversidade. Possui uma camada de *Aplicação Cliente* responsável por reunir, processar e exibir os dados ao usuário. Essa camada irá acessar vários *Serviços Web de Consultas*, incluindo um serviço de consulta de imagens baseado em conteúdo, um serviço de consultas geográficas, um serviço de ontologias e um serviço de metadados. Esses serviços acessam *Coleções de Dados*. Esta dissertação trata do *Serviço de Ontologias*. O alvo da implementação é uma ferramenta geral, que possa ser utilizada nas mais diversas aplicações que necessitem de contextualização semântica através de ontologias. A Figura 1 ilustra a arquitetura proposta para o Serviço de Ontologias, composta basicamente por uma *Interface* de comunicação, as *Funcionalidades* providas pelo serviço e uma *Base de Ontologias*.

A *Interface* delimita as entradas possíveis e funcionalidades providas. As entradas

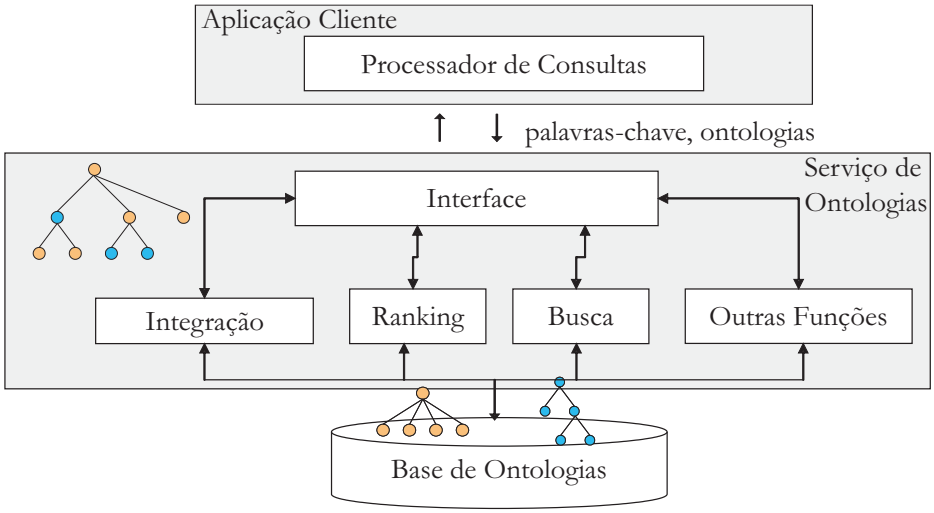

Figura 1. Arquitetura Proposta

previstas são palavras-chave e ontologias. As funcionalidades, na maioria dos casos, trabalharão conjuntamente. A função *Busca*, que busca as palavras-chave na base de ontologias, está intimamente relacionada à função *Ranking*, que deverá selecionar quais ontologias são relevantes no contexto pesquisado.

O módulo *Integração* deverá integrar as ontologias selecionadas pelo *Ranking* na busca por termos em ontologias diferentes. Ele irá gerar uma nova ontologia a partir da qual será extraído o resultado. O serviço será usado em consultas multimodais sobre dados de biodiversidade. O módulo deve contextualizar termos relacionados aos diferentes dados em uma mesma ontologia. Os pontos em aberto na *Integração* incluem uma análise das necessidades das ontologias de biodiversidade que justifiquem a escolha de uma estratégia associada a algumas das técnicas de integração descritas na seção 2.

A metodologia adotada irá se basear no levantamento dos parâmetros de sistemas de biodiversidade no que tange aspectos de ontologias. A partir deles, pretende-se especificar as solicitações a um serviço de ontologias - camadas e parâmetros, e suas principais funcionalidades. O serviço será testado com ontologias de biodiversidade, construídas com auxilio dos usuários-alvo, pesquisadores em biodiversidade do Instituto de Biologia da UNICAMP que fazem parte do projeto *WeBios*.

4. Trabalhos e Iniciativas Similares

Encontrar manualmente similaridades entre ontologias é uma tarefa tediosa e propensa a erros. Isso ocasionou o surgimento de várias ferramentas que encontram similaridades entre ontologias de forma semi-automática. Esta seção descreve algumas das abordagens implementadas na literatura. A Figura 2 apresenta uma análise comparativa das características mais relevantes relacionadas às ferramentas de integração de ontologias citadas.
GLUE [Doan et al. 2002]: ferramenta que auxilia o processo de mapeamento de ontologias, empregando estratégias de aprendizado para criar mapeamentos semânticos de forma semi-automática. Considera as taxonomias como componentes centrais das ontologias dadas, buscando encontrar correspondências um-para-um entre seus nós.
Chimaera [McGuinness et al. 2000]: ambiente *Web* para união de ontologias. A ferramenta sugere, durante o processo de união de duas ontologias, os potenciais candidatos a

união, baseando-se em um conjunto de propriedades (como similaridades de nomes), que são localizados nas hierarquias correspondentes. O usuário pode escolher uni-los ou não a partir de navegação nas hierarquias.

ODEMerge [Ramos 2001]: ferramenta cliente-servidor via *Web* para união de ontologias, integrada no WebODE, uma plataforma de software para construir ontologias. O processamento é feito juntamente com informações de tabelas de sinônimos e hiperônimos, gerando uma nova ontologia como resultado.

PROMPT [Noy and Musen 2000]: ferramenta semi-automática para guiar o alinhamento e a união de ontologias. Direciona o usuário identificando possíveis pontos de integração, e fazendo sugestões considerando quais operações devem ser feitas posteriormente a uma modificação, quais conflitos precisam ser resolvidos e como esses conflitos podem ser resolvidos. O AnchorPrompt é uma extensão do PROMPT que compara a estrutura de grafo das ontologias, analisando os caminhos nos subgrafos e determinando quais classes freqüentemente aparecem em posições ou em caminhos similares.

CATO (Componente para Alinhamento Taxonômico de Ontologias) [Felicíssimo 2004]: protótipo de validação de uma estratégia de alinhamento de ontologias. A estratégia adotada foca a taxonomia como o componente central das ontologias, considerando inicialmente apenas os conceitos com relacionamentos de especialização do tipo "é-um" identificados nas ontologias de entrada.

Ferramenta	**GLUE**	**Chimaera**	**ODEMerge**	**PROMPT**	**CATO**
Tipo de Integração	Mapeamento	União	União	Alinhamento e União	Alinhamento
Técnicas Aplicadas	Análise taxonômica, técnicas em string (similaridade) e restrições	Análise taxonômica, técnicas em string (similaridade)	Recursos lingüísticos (sinônimos, hiperônimos)	Análise taxonômica, técnicas em string, restrições de tipo, técnicas de grafos (Anchor-PROMPT)	Análise taxonômica, recursos lingüísticos (sinônimos)
Ambiente de Apoio	Nenhum	Ontolingua	WebODE	Protégé	Nenhum
Formato das Ontologias	Formato Taxonomia	Ontolingua, Protégé, XOL	XML, RDF(S), DAML+OIL	Protégé, RDFS, XML Schema	OWL
Tipo de Processamento	Semi-automático	Iterativo	Automático	Semi-automático	Automático
Adequabilidade Domínio das Ontologias	Mesmo domínio	Domínios similares ou sobrepostos	Domínios similares ou sobrepostos	Domínios complementares	Domínio complementares similares e sobreposição

Figura 2. Tabela Comparativa das Ferramentas de Integração de Ontologias

5. Resultados Esperados e Aplicabilidade das Contribuições

Este projeto atende a demanda por sistemas caracterizados pela multiplicidade de usuários e visões do mundo, o que gera uma grande variedade de nomes e conceitos, fonte de heterogeneidade. Tais sistemas de informação cada vez mais exigem cooperação de usuários de diferentes ramos da ciência, o que traz à tona questões de semântica de conceitos. As principais contribuições esperadas deste projeto são: (1) estudo das ferramentas de manipulação de ontologias disponíveis atualmente; (2) levantamento dos problemas inerentes à construção de ontologias e possibilidades de integração; (3) levantamento das

principais ontologias existentes para a área de biodiversidade; (4) especificação das necessidades de um serviço de ontologias; e (5) especificação e implementação de um servidor de ontologias para o contexto de sistemas de biodiversidade.

A validação do trabalho será realizada através do projeto *WeBios* [web 2005]. Serão realizados testes a partir dos dados fornecidos pelo Instituto de Biologia. Prevê-se a criação de ontologias na área de biodiversidade para contextualizar esses dados.

Agradecimento: Apoio financeiro recebido da CAPES, FAPESP (processo 05/57424-0), e *Microsoft Research* financiadora do projeto *WeBios*.

Referências

(started 2005). Webios: Web service multimodal tools for strategic biodiversity research, assessment and monitoring. http://www.lis.ic.unicamp.br/projects/webios.

Abels, S., Haak, L., and Hahn, A. (2005). Identification of common methods used for ontology integration tasks. In *IHIS '05: First International Workshop on Interoperability of Heterogeneous Information Systems*, pages 75–78. ACM Press.

Doan, A., Madhavan, J., Domingos, P., and Halevy, A. (2002). Learning to map between ontologies on the semantic web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 662–673. ACM Press.

Ehrig, M., Bruijn, J., Manov, D., and Martín-Recuerda, F. (2004). State-of-the-art survey on ontology merging and aligning. SEKT Deliverable 4.2.1, DERI Innsbruck.

Felicíssimo, C. H. (2004). Interoperabilidade semântica na web: Uma estratégia para o alinhamento taxonômico de ontologias. Master's thesis, PUC-Rio de Janeiro.

Kalfoglou, Y. and Schorlemmer, M. (2003). Ontology mapping: the state of the art. *Knowledge Engineering Review*, 18(1):1–31.

McGuinness, D. L., Fikes, R., Rice, J., and Wilder, S. (2000). The chimaera ontology environment. In *Proc of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 1123–1124.

Noy, N. F. and Musen, M. A. (2000). PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Seventeenth International Joint Conference on Artificial IntelligenceAAAI/IAAI*, pages 450–455.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350.

Ramos, J. A. (2001). Mezcla automática de ontologías y catálogos electrónicos. Final Year Project. Facultad de Informática de la Universidad Politécnica de Madrid. Spain.

Shvaiko, P. and Euzenat, J. (2004). A survey of schema-based matching approaches. Technical Report DIT-04-087, University of Trento.

Torres, R. S. (2004). *Ambiente de Gerenciamento de Imagens e Dados Espaciais para Desenvolvimento de Aplicações em Biodiversidade*. PhD thesis, IC-UNICAMP.