Um Servidor de Ontologias para 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@students.ic.unicamp.br, cmbm@ic.unicamp.br

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. This research is directly connected with the first challenge (management of large multimedia data volumes), and provides support to research in challenge 2 (computational modeling in complex systems).*

Resumo. *A pesquisa em biodiversidade requer correlacionar dados sobre seres vivos e seus habitats, integrando desde relacionamentos espaciais a especificações de domínio, freqüentemente 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. A pesquisa está diretamente relacionada com o primeiro desafio (gerenciamento de grandes volumes de dados multimídias) e provê subsídios a pesquisa no segundo desafio (modelagem computacional em sistemas complexos).*

1. Introdução

O primeiro grande desafio se preocupa com o gerenciamento de grandes volumes de dados heterogêneos e distribuídos, característica do ambiente Web atual. O segundo grande desafio trata de modelagem computacional de fenômenos e interações complexos. A pesquisa em *e-Science* é um dos domínios recentes em que ambos os desafios se entrecruzam - exige solução para problemas de modelagem, ao mesmo tempo em que se depara com questões de gerenciamento de dados científicos, produzidos em larga escala e em tempo real. Uma das perspectivas de longo prazo, inclusive, é que soluções para um problema dentro de um desafio contribuam para resolver problemas de outro.

A pesquisa em biodiversidade é um um exemplo típico desse cruzamento. Assunto de grande destaque, requer correlacionar dados sobre seres vivos e seus habitats. Estima-se que, somente no Brasil, haja cerca de 200.000 espécies reconhecidas de um total de

1.900.000 espécies existentes. Dada esta variedade, é preciso soluções para a coleta, o processamento e a descrição de dados de biodiversidade.

Sistemas de Informação de Biodiversidade [Torres et al. 2006] visam solucionar este problema, de forma a permitir que os pesquisadores analisem características das espécies e suas interações. Tais sistemas exigem manipulação de um grande número de tipos de dados, cobrindo desde a recuperação de informações textuais, como descrições taxonômicas e literais, à combinação de informações taxonômicas com a distribuição espacial de uma ou mais espécies.

Um dos desafios para o desenvolvimento de tais sistemas é a manipulação e análise, de forma integrada, de dados heterogêneos provenientes de coleções de biodiversidade de diferentes pesquisadores. Além dos problemas intrínsecos à heterogeneidade, volume e distribuição, fatores externos como a classificação de espécies e modelos ecológicos também mudam, refletindo a evolução do conhecimento científico no mundo real. A ausência da padronização na representação semântica dos dados torna o problema da interoperabilidade ainda mais complexo. O uso de ontologias tem sido apontado como solução para alguns desses problemas.

Ontologias são descrições de um modelo abstrato de termos, relacionados entre si [Gruber 1995]. Modelam uma parte da realidade, suas entidades, relações taxonômicas, não-taxonômicas e restrições, visando definir um entendimento comum sobre um domínio. Ontologias têm adquirido importância em aplicações industriais e acadêmicas, sendo usadas em buscas semânticas, especificações de restrições, interoperabilidade entre aplicações, dentre outras.

O WeBios [WeBios 2005] é um projeto conduzido por pesquisadores dos Institutos de Computação e de Biologia da UNICAMP que se encaixa neste contexto. Seu objetivo é prover um sistema que permita consultas exploratórias multimodais sobre fontes distribuídas e heterogêneas de dados biodiversidade. As fontes consideradas incluem dados textuais de espécies, imagens, dados geográficos, ontologias e anotações, acessadas via serviços Web.

Este artigo descreve um novo tipo de mecanismo - um Serviço de Ontologias - que está sendo especificado e implementado no IC-Unicamp para o WeBios. O serviço visa prover acesso, manipulação, análise e integração de ontologias. Com isto, espera-se auxiliar a solução de problemas de heterogeneidade e associar mais semântica às operações de sistemas que manipulem grande volume de informações distribuídas. A pesquisa está diretamente relacionada com o primeiro desafio e provê subsídios a pesquisa no segundo desafio (modelagem computacional em sistemas complexos).

A própria especificação deste serviço já representa uma contribuição para o primeiro Grande Desafio: não apenas não existe na literatura menção de serviços deste tipo, como também a intenção final é ter uma ferramenta de usabilidade geral, que possa ser utilizada nas mais diversas aplicações que manipulem grandes volumes de dados heterogêneos na Web e que necessitem de contextualização semântica através de ontologias.

O restante deste texto está organizado da seguinte forma. A seção 2 apresenta a arquitetura do sistema WeBios. A seção 3 descreve os principais trabalhos relacionados à manipulação de ontologias e a seção 4 apresenta a especificação do Serviço de Ontologias proposto. As conclusões e perspectivas a longo prazo são apresentados na seção 5.

2. O Projeto WeBios

O WeBios [WeBios 2005] é um sistema centrado em serviços Web que suporta consultas exploratórias multimodais sobre fontes de dados heterogêneas de biodiversidade, para cientistas que trabalham com questões ambientais e de biodiversidade. As fontes de dados incluem imagens (fotos de seres vivos ou seus habitats), dados geográficos (mapas de regiões com ocorrência de espécies), ontologias e metadados específicos do domínio (descrições do habitat e ecossistema). Visa permitir que os cientistas incrementem seus conhecimentos sobre espécies, suas interações e seus habitats e correlações entre eles.

O principal diferencial em relação aos demais Sistemas de Biodiversidade existentes é permitir a combinação de predicados baseados em conteúdo, espaciais e textuais tradicionais (de ontologias e metadados) em uma mesma consulta [Torres et al. 2006]. Os sistemas disponíveis publicamente não atacam estas questões simultaneamente; eles se concentram ou apenas em dados de imagem ou apenas em dados espaciais.

A Figura 1 ilustra a arquitetura do WeBios, composta de três camadas principais: a *Camada de Armazenamento*, os *Serviços de Suporte* e os *Serviços Avançados*. O módulo *Aplicação Cliente* é responsável por reunir, processar e exibir os dados ao usuário.

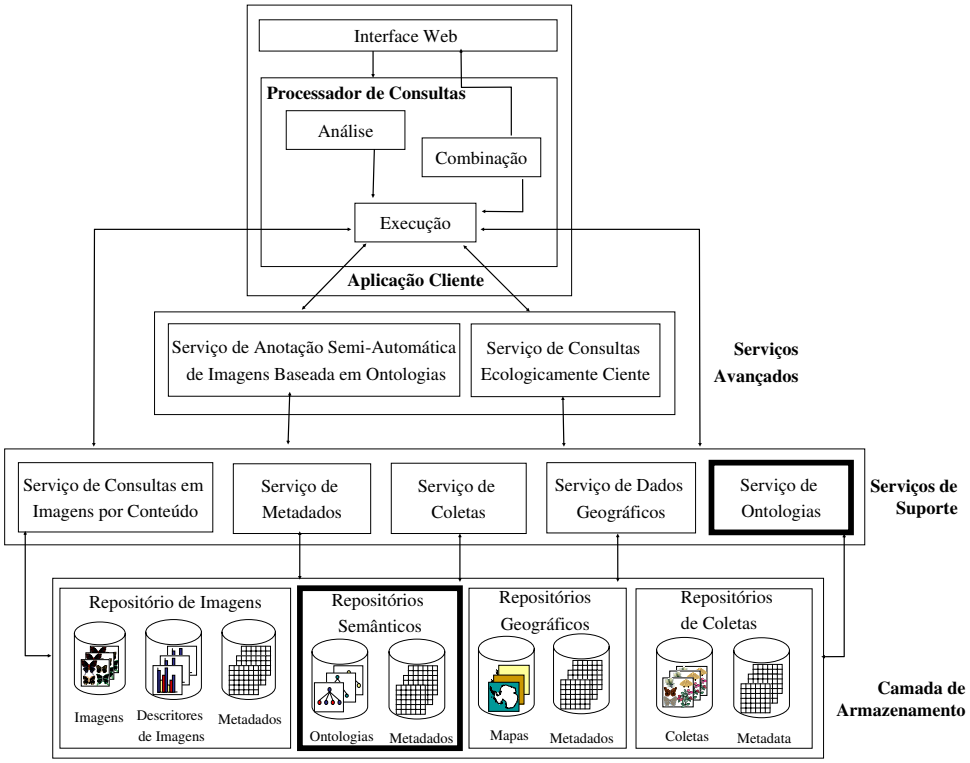

Figura 1. Arquitetura do Sistema WeBios

O sistema possui serviços de suporte para consulta em imagens por conteúdo e gerência de metadados, dados de coletas, dados geográficos e ontologias. Além disso, o sistema inclui serviços avançados que visam combinar as funcionalidades de dois ou mais serviços de suporte provendo funcionalidades mais complexas. O serviço de consultas ecologicamente ciente permite processar relações ecológicas descritas em ontologias

e dados geográficos enquanto o serviço de anotação de imagens utiliza conceitos de ontologias para anotação e busca de imagens.

O escopo deste artigo se refere ao Serviço de Ontologias, que utiliza Repositórios Semânticos, ambos destacados na figura 1. Sua especificação e implementação são baseadas no paradigma de Serviços Web, tendência mundial para garantir interoperabilidade e independência de plataforma na Web.

3. Trabalhos Relacionados

3.1. Visão Geral

Ontologias vêm sendo adotadas para apoio à interoperabilidade de aplicações. Em muitos domínios, é comum que os especialistas desenvolvam suas próprias ontologias, adequadas à suas aplicações. Em outros casos, é possível reutilizar ontologias existentes. Vários repositórios têm sido criados recentemente para armazenar e compartilhar ontologias, como DOME, OntoServer, Ontaria, Swoogle [Ding et al. 2004].

A exploração da semântica dos dados nessas aplicações depende de mecanismos que permitam a manipulação adequada dessas ontologias. Atualmente, existem vários *frameworks* para o desenvolvimento de aplicações que manipulem ontologias, como Jena [Carroll et al. 2004], SNOBASE [Lee 2003] e SOFA[1].

Em muitos casos, o uso de tais *frameworks* torna estática a exploração da semântica nas aplicações. Vários servidores de ontologias foram propostos para suprir a necessidade de gerenciamento dinâmico de ontologias [Li et al. 2003, Suguri et al. 2001, Duke and Patel 2003]. De forma similar aos *frameworks*, os servidores também provêem a funcionalidade de consulta de ontologias e, em alguns casos, fornecem também máquinas de inferência. Alguns servidores provêem acesso às ontologias através de suas URIs *(Unified Resource Identifiers)* e outros armazenam as ontologias em repositórios locais, facilitando o gerenciamento e controle de versões.

Em cenários distribuídos, entretanto, as funcionalidades providas por esses serviços ainda não são suficientes, já que só permitem acesso a uma ontologia por vez. É necessária a manipulação de várias ontologias de forma conjunta, seja na comparação de suas estruturas em relação a um conceito ou na descoberta de mapeamento entre seus conceitos. Muitas aplicações necessitam de apenas parte do contexto descrito por ontologias de outras aplicações e, em alguns casos, a comparação de versões de uma ontologia é necessária para contextualização de dados. Não existe uma solução que engloba todas as funcionalidades necessárias e, além disto, algumas delas não estão disponíveis na Web. Nosso serviço visa preencher tal lacuna.

3.2. Ranking de Ontologias

Mecanismos de *ranking* visam determinar as ontologias potencialmente relevantes para um domínio, baseando-se em diferentes critérios. Alguns mecanismos adotam a análise de *links* e referências entre ontologias. Este método baseado em popularidade, similar ao *Page-Ranking*, é utilizado pelas ferramentas Swoogle [Ding et al. 2004] e OntoKhoj [Patel et al. 2003]. Entretanto, a baixa conectividade de grande parte das ontologias disponíveis restringe a abrangência do método. Outras técnicas de *ranking*, como

[1]http://sofa.dev.java.net

a utilizada pela ferramenta AkTiveRank [Alani et al. 2006], analisam as estruturas internas da ontologia. Essa abordagem baseia-se em métricas que avaliam como a ontologia representa os conceitos de interesse, considerando suas taxonomias e propriedades.

3.3. Detecção de Modificações em Ontologias

A detecção de modificações compara, na maioria das vezes, duas versões de uma mesma ontologia, identificando as diferenças existentes entre elas – como nomes de classes, ou hierarquias taxonômicas. Existem várias ferramentas que tratam da detecção de modificações em ontologias. Algumas abordagens são concentradas na evolução de ontologias e na propagação de modificações. A ferramenta OntoDiff [Tury and Bieliková 2006], por exemplo, detecta modificações entre duas versões de uma ontologia, identificando termos adicionados, removidos e modificados.

3.4. Visões de Ontologias

Especificar visões de uma ontologia permite limitar um subconjunto do domínio relevante para a aplicação. A visão consiste em extrair partes da ontologia, como uma sub-ontologia. Algumas abordagens utilizam linguagens de consulta para ontologias para definir visões [Volz et al. 2003]. Uma proposta de extração automática de visões é apresentada em [Alani et al. 2005]. Nessa proposta, as consultas realizadas às ontologias são analisadas e a visão é uma sub-ontologia que possui os elementos mais freqüentemente requisitados pelas consultas da aplicação.

A proposta de [E. Jiménez 2005] define uma linguagem para a definição de visões centradas em um conceito inicial, contendo operadores para a seleção de conceitos, propriedades e instâncias para ontologias em RDF(S). De forma similar, [Noy and Musen 2004] apresenta o conceito de *View Traversal* para definir quais elementos da ontologia a visão deve conter.

3.5. Integração de Ontologias

Em sistemas distribuídos e abertos não é possível evitar a heterogeneidade dos dados apenas com o uso de ontologias. Grupos de pesquisadores podem diferir em interesses, focos de pesquisa, em ferramentas usadas ou ainda manipular o conhecimento em diferentes níveis de detalhe. A integração de ontologias é uma solução cada vez mais comum para definir as relações existentes entre as representações heterogêneas. O processo de integração permite, por exemplo, criar expressões compatíveis entre ontologia diferentes.

As abordagens para a integração de ontologias incluem [Kalfoglou and Schorlemmer 2003, Bruijn et al. 2004]: (1) mapeamento: identifica entidades idênticas entre todos os conceitos de duas ontologias; (2) união: realiza a fusão de ontologias de acordo com os mapeamentos existentes, resultando na construção de uma nova ontologia; e (3) alinhamento: produz um conjunto de mapeamentos entre ontologias fazendo com que a conceitualização e o vocabulário se emparelham em algumas partes, porém mantendo integralmente as ontologias originais.

A identificação das similaridades entre ontologias é, em geral, baseada na análise e no reconhecimento de partes que “casam” umas com as outras. Este casamento pode ser a identificação de partes idênticas (equivalências) ou de elementos de relacionamentos (ex. parte-de, é-um). O casamento pode considerar pares de elemen-

tos isolados, ou analisar como tais elementos aparecem na estrutura de uma ontologia [Shvaiko and Euzenat 2004].

Há algumas ferramentas na literatura para encontrar similaridade entre ontologias, realizando diferentes técnicas de integração [Doan et al. 2002, McGuinness et al. 2000, Ramos 2001, Noy and Musen 2000, Felicíssimo 2004]. A maioria delas são integradas a ambientes de manipulação de ontologias, como o Protégé (PROMPT) e o Ontolingua (Chimaera). As ferramentas de processamento automático geram mapeamentos incorretos em algum casos, enquanto as de processamento interativo muitas vezes sobrecarregam o usuário na verificação de todos os mapeamentos encontrados. Em geral, as ferramentas que combinam técnicas para analisar os elementos tanto isoladamente como na estrutura das ontologias apresentam melhores resultados de integração.

4. Especificação do Serviço de Ontologias

O objetivo do Serviço de Ontologias é prover acesso, manipulação, análise e integração de ontologias. O paradigma proposto para a implementação é acesso via Serviços Web (SW). A Figura 2 ilustra a arquitetura do Serviço de Ontologias, composta por duas camadas: Repositórios Semânticos (Seção 4.1) e Operações (Seção 4.2). Os componentes das camadas são encapsulados por serviços Web. Ambas camadas podem ser acessadas por aplicações cliente, de acordo com as funcionalidades solicitadas.

O gerenciamento das ontologias e seus metadados é provido pela camada de Repositórios Semânticos, enquanto a camada de Operações é responsável pelas funcionalidades avançadas do serviço. Cada Repositório Semântico é composto por ontologias e metadados associados, sendo acessados por um serviço Web específico. A figura mostra que, adicionalmente, o usuário pode explicitar Repositórios de Ontologias para efetuar certas operações, desde que estes possam ser acessados via serviços Web.

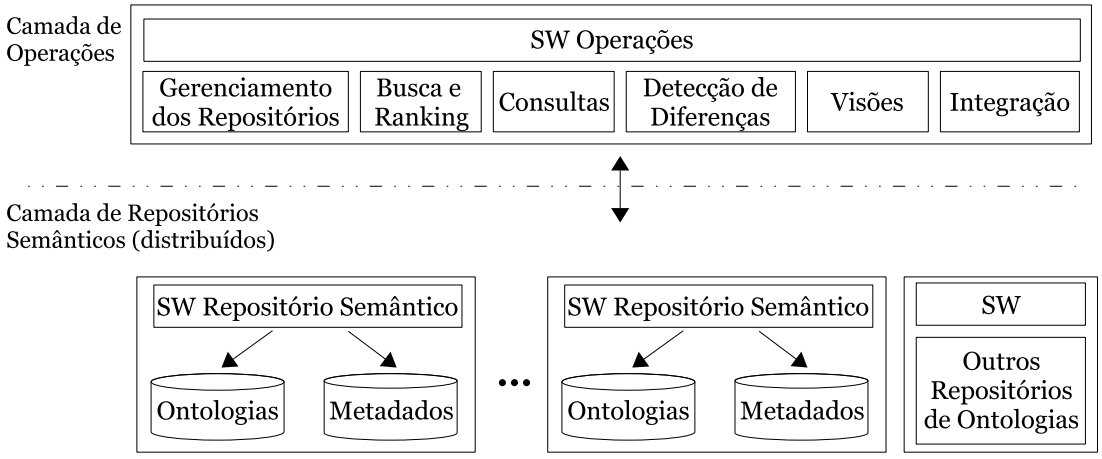

Figura 2. Arquitetura do Serviço de Ontologias

4.1. Camada de Repositórios Semânticos

Esta camada é composta pelos repositórios distribuídos de ontologias e de metadados. A linguagem de representação de ontologias adotada pelo Serviço de Ontologias é a OWL *(Web Ontology Language)* [Antoniou and van Harmelen 2003], recomendada pelo

consórcio W3C. OWL é baseada em XML, como a maioria das linguagens para ontologias. A escolha dessa linguagem se deve ao fato de não se saber, previamente, o nível de detalhamento das ontologias que serão manipuladas pelo serviço. As ontologias poderão ser extraídas de repositórios Web ou criadas pelos próprios usuários. Optou-se, portanto, por uma linguagem padrão que permitisse representar diferentes níveis de expressividade.

As ontologias manipuladas por esta camada relacionam-se aos tipos de dados gerenciados pelo sistema WeBios, envolvendo os conceitos: características geográficas, biológicas e associações entre metadados. Ontologias que descrevem características geográficas são relacionadas ao Serviço Geográfico, e as correspondências semânticas entre metadados são providas pelo Serviço de Metadados (vide figura 1). Ontologias com informações biológicas possuem, dentre outros, descrições de taxonomias e filogenias, evolução e morfologia de espécies e relacionamentos ecológicos entre espécies.

Atualmente, a maioria das ontologias são difundidas na Web sem informações adicionais. Essa deficiência afeta seriamente seu compartilhamento e reuso, pois a ausência de metadados faz com que potenciais usuários não consigam encontrar e identificar suas ontologias de interesse. Para evitar esse tipo de problema, a camada de repositórios é composta de Repositórios Semânticos, que são repositórios distribuídos contendo, além das ontologias, estruturas de metadados das ontologias. O padrão adotado é o OMV *(Ontology Metadata Vocabulary)*, uma iniciativa de padronização de metadados para ontologias [Hartmann et al. 2005].

O esquema de metadados OMV é formalizado com uma ontologia, descrita na linguagem RDF *(Resource Description Language)*. Os elementos especificados pela OMV seguem a seguinte classificação:

- **Geral:** elementos que fornecem informações gerais sobre a ontologia;
- **Disponibilidade:** localização da ontologia na Web (URI ou URL);
- **Aplicabilidade:** intensão de uso ou escopo da ontologia;
- **Formato:** representação física, incluindo a linguagem de representação em que a ontologia foi formalizada;
- **Origem:** organizações e colaboradores na criação da ontologia;
- **Relacionamentos:** informações sobre relacionamentos com outras ontologias. Inclui referências a outras versões ou importações de outras ontologias;
- **Estatísticas:** métricas fundamentadas na topologia de grafos da ontologia.

A Figura 3 ilustra parcialmente uma ontologia sobre interações entre insetos e plantas, criada para o WeBios com apoio dos biólogos do IB-UNICAMP. A figura mostra, por exemplo, um inseto da família *Agromyzidae* que possui uma interação com uma planta da espécie *Chromolaena odorata*. Essa interação, como descrita na ontologia, classifica o inseto como um herbívoro endofágico brocador.

Operações de consulta e atualização nos repositórios desta camada são feitas a partir da Camada de Operações. Os repositórios são mantidos por usuários especialistas. A Camada de Operações solicita consultas e atualizações à Camada de Repositórios por meio do envio de mensagens SOAP, contendo arquivos OWL e RDF anexados.

4.2. Camada de Operações

A camada de operações é responsável pelas funcionalidades avançadas do serviço. Essas funcionalidades foram definidas a partir do levantamento de requisitos junto aos

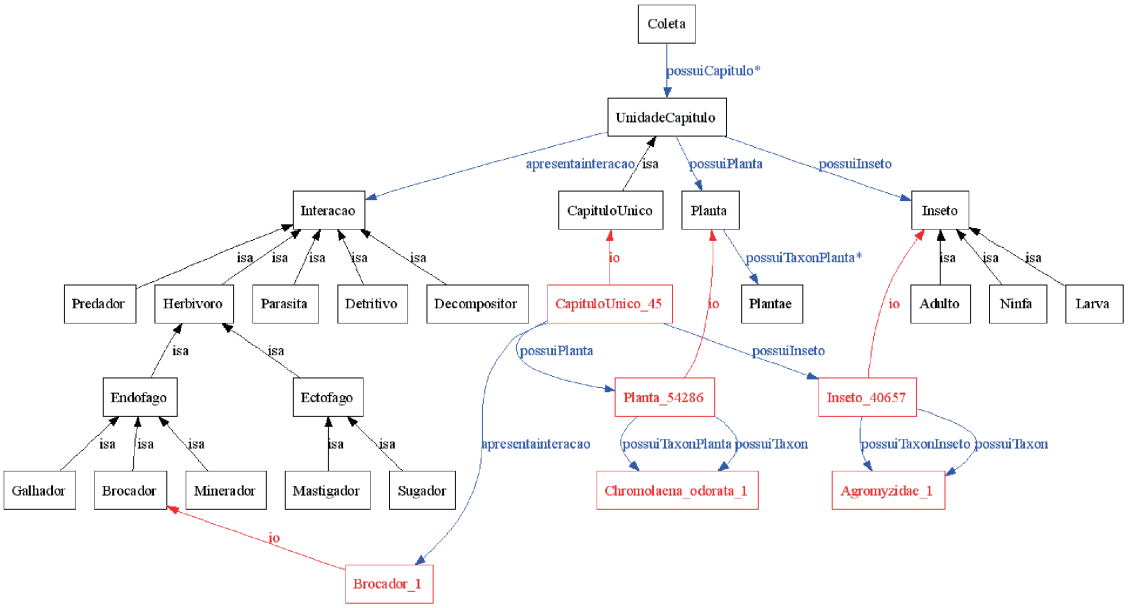

Figura 3. Exemplo de Ontologia Biológica

biólogos parceiros do projeto WeBios e também de análise de sistemas Web de biodiversidade. As invocações aos módulos recebem os parâmetros de conexão com os repositórios semânticos e a recuperação de ontologias é realizada com base em seus identificadores – pares URL-repositório, ID-ontologia. Novas ontologias geradas pela execução de algum módulo desta camada são armazenadas nos Repositórios Semânticos.

Como já mencionado, este trabalho aborda questões relacionadas ao primeiro Grande Desafio: gestão da informação em grandes volumes de dados multimídia distribuídos. A camada de operações é o núcleo da proposta, pois seus módulos embutem a lógica necessária para diferentes tipos de manipulação de dados com tais características. Além disso, como o domínio alvo é biodiversidade, as funções servem de apoio a necessidades do segundo desafio (modelagem computacional), no que tange fenômenos naturais.

Módulo de Gerenciamento de Repositórios

Este módulo é responsável pelo gerenciamento de ontologias e metadados da camada de repositórios. Este gerenciamento inclui as funções de armazenamento, substituição e eliminação. Cada ontologia armazenada possui um identificador único no seu repositório, que é utilizado para seu acesso. Cada conjunto de metadados é associado a uma ontologia, sendo inserido no mesmo repositório da ontologia correspondente.

Módulo de Busca e *Ranking*

Realiza a busca de ontologias tanto nos repositórios da camada inferior quanto em outros repositórios Web designados pelo usuário. São retornadas ontologias que possuem classes ou instâncias cujos nomes “casam” (exata ou parcialmente) com os termos fornecidos. As ontologias retornadas pela busca são armazenadas no repositório da camada inferior, para acelerar buscas posteriores.

O módulo permite dois tipos de operações: a busca com *ranking* e sem *ranking*,

sendo que a última retorna uma única ontologia. A invocação de uma busca com *ranking* visa retornar um conjunto de ontologias e possui a forma:

$$BuscaRank(\{termos\}, \{pesos\}, \{repWeb\}),$$

em que $\{termos\}$ representa o conjunto de termos da busca e $\{pesos\}$ representa o conjunto de pesos para as métricas de *ranking*, e $\{repWeb\}$ designa repositórios indicados pelo usuário, tanto Semânticos quanto repositórios adicionais na Web. O repositório de ontologias Swoogle[2] [Ding et al. 2004] é *default* utilizado na busca Web.

O *ranking* é baseado na análise de estruturas internas de cada ontologia retornada, suas taxonomias e propriedades, combinando 4 métricas para avaliar como a ontologia representa os conceitos de interesse, combinadas de acordo com pesos forncecidos. Essas métricas analisam, por exemplo, a densidade e a centralidade das classes identificadas pelos termos buscados e a similaridade semântica entre essas classes. O retorno dessa função é uma lista ordenada de identificadores de ontologias, seguindo as métricas de *ranking*. A busca sem *ranking* retorna uma ontologia contendo hierarquias taxonômicas de espécies em OWL, e a invocação desta busca possui a forma:

$$Busca(taxon, \{diretivas\}, \{repWeb\}),$$

em que $taxon$ representa o nome do táxon buscado e $\{diretivas\}$ representa o conjunto de diretivas utilizadas na recuperação da hierarquia: antecessores, descendentes e táxons de mesmo nível ("irmãos"). O portal do projeto Spire[3] [Parr et al. 2006] é utilizado como repositório Web *default*.

Módulo de Consultas em Ontologias

A invocação do módulo de consulta possui a forma:

$$Consulta(idOnto, linguagem, stringConsulta, formato),$$

em que o $idOnto$ representa o identificador da ontologia a ser consultada (em um Repositório Semântico), $linguagem$ corresponde à linguagem utilizada na especificação da consulta, contida na $stringConsulta$. O módulo permite o uso das linguagens RDQL [Seaborne 2003] e SPARQL [Prud'hommeaux and Seaborne 2006]. O campo $formato$ define o formato de saída para consultas SPARQL, que pode ser textual (partes das triplas) ou estruturado em arquivos XML. O resultado de consultas RDQL é sempre textual.

Ambas linguagens executam consultas sobre modelos ontológicos RDF, podendo também ser utilizadas em modelos OWL uma vez que estes são descritos sobre RDF. O processamento de consultas em ontologias considera estas como um conjunto de triplas (Sujeito-Propriedade-Valor). As consultas consistem basicamente em buscar quais dessas triplas satisfazem os predicados.

O processamento semântico das consultas varia de acordo com a linguagem de representação de ontologias utilizada. Na linguagem OWL, por exemplo, o uso de máquinas de inferência pode fornecer resultados mais expressivos, utilizando as relações de equivalência de classes e transitividade de propriedades. Por esse motivo, o módulo constrói um modelo inferido da ontologia antes de executar a consulta solicitada.

[2]http://swoogle.umbc.edu/

[3]http://spire.umbc.edu/ont/ethan.php

Módulo de Detecção de Diferenças entre Ontologias
O módulo de detecção estabelece uma comparação entre duas ontologias, tanto em suas estruturas quanto seus conteúdos. A comparação das estruturas OWL considera apenas a hierarquia das classes e suas instâncias. A invocação do módulo possui a forma:

$$Diferenca(idOnto1, idOnto2),$$

em que $idOnto1$ e $idOnto2$ são identificadores de ontologias. O resultado deste módulo é um arquivo XML contendo as diferenças encontradas.

Esta funcionalidade possui aplicações interessantes em sistemas de biodiversidade, especificamente para ontologias biológicas taxonômicas e evolutivas. A classificação de espécies pode sofrer modificações ao longo do tempo, além de haver discordância entre autores sobre a identificação de algumas espécies. Uma operação de comparação permite ao usuário, por exemplo, detectar quando diferentes modelos taxonômicos foram utilizados em algum estudo.

Módulo de Visões de Ontologias
A construção de visões de ontologias é baseada na abordagem de conceito central [Noy and Musen 2004] descrita na seção 3.4, estendida para a manipulação dos axiomas que definem as classes. A invocação deste módulo possui a forma:

$$Visao(idOnto, conceito, \{diretivas\}, repSem),$$

em que $idOnto$ é o identificador da ontologia, $conceito$ representa o conceito central da visão e $\{diretivas\}$ representa o conjunto de diretivas, contendo os elementos que devem ser abrangidos pela visão. Esses elementos podem ser: instâncias, axiomas, e propriedades com profundidades associadas. As propriedades podem incluir subclasses, superclasses, propriedades entre classes ou atributos. Uma invocação sem diretivas produz uma visão contendo todos os elementos da ontologia relacionados com o conceito central, com profundidades "infinitas".

A visão construída como resultado é uma sub-ontologia da ontologia identificada por $idOnto$, que é armazenada no Repositório designado em $repSem$. Além da classe central, as classes selecionadas para fazerem parte da visão incluem as que se relacionam com esta através de propriedades ou axiomas.

Módulo de Integração de Ontologias
Dentre as abordagens de integração - mapeamento, união e integração, descritas na seção 3.5 - a que mais se adequa às necessidades do serviço é o alinhamento. Isto se deve ao fato de que, em biodiversidade, as ontologias manipuladas pertencerão, na maioria dos casos, a domínios complementares ou sobrepostos. Além disso, a união ocasionaria a junção de termos equivalentes privilegiando a terminologia de uma das ontologias, o que não é desejável. Esta função é importante para contextualizar termos relacionados a dados de diferentes fontes em consultas no WeBios. A invocação deste módulo possui a forma:

$$Alinhamento(idOnto1, idOnto2, repSem),$$

em que $idOnto1$ e $idOnto2$ são identificadores de ontologias. O módulo combina técnicas que analisam os elementos tanto isoladamente como na estrutura das ontologias na descoberta dos mapeamentos. A nova ontologia construída é armazenada no Repositório Semântico $repSem$ e seu identificador é retornado pela operação.

4.3. Estado Atual da Implementação

As ontologias utilizadas para teste do serviço estão sendo criadas com a ferramenta Protégé [Gennari et al. 2003]. A ferramenta possui um *plugin* para a construção de ontologias OWL, com uma interface gráfica para a definição de classes, propriedades, instâncias e restrições. O protótipo do Serviço de Ontologias está sendo implementado na linguagem Java, e a navegação no conteúdo das ontologias está sendo provido pelo *framework* Jena versão 2.4 [Carroll et al. 2004].

Jena possui uma API para manipulação de ontologias em OWL, permite seu armazenamento em bancos de dados relacionais, suporta as linguagens de consulta e possui mecanismos de inferência baseadas em regras. Ressaltamos que, embora Jena 2.5 já esteja disponível, esta nova versão não suporta a linguagem de consulta RDQL. Os desenvolvedores de Jena seguem as recomendações da W3C e consideram que RDQL é obsoleta, havendo assim removido suas bibliotecas do *framework*. Apesar desta linguagem não constar nas recomendações da W3C, ela ainda é muito utilizada. Por esta razão, nosso Serviço de Ontologias utiliza a versão 2.4 do *framework*.

No módulo de integração, a descoberta dos mapeamentos entre as ontologias utiliza a comparação da similaridade das *strings*, através da distância de edição, usando a WordNet [Kong et al. 2005] como dicionário de sinônimos para termos gerais. Além disso, os candidatos ao mapeamento são analisados estruturalmente, de acordo com suas propriedades, atributos e axiomas, na tentativa de evitar casos em que termos identificados por um mesmo nome, em contextos diferentes, sejam identificados como similares. Os alinhamentos encontrados entre classes são representados em OWL com <owl:equivalentClass> e entre instâncias com <owl:sameAs>. Com isso, ao utilizar mecanismos de inferência, as instâncias de cada uma das classes alinhadas abrangem também as instâncias da classe equivalente.

As técnicas de alinhamento de ontologias precisam ser adaptadas no caso de termos contendo táxons de espécies. Os nomes científicos dos táxons, muitas vezes, se diferenciam apenas terminação. As *strings* "Asterales" e "Asteraceae", por exemplo, possuem alta similaridade em algoritmos de distância de edição, embora tratem de táxons diferentes: o sufixo *-ales* determina ordem, enquanto o sufixo *-aceae* indica família. Ontologias com táxons ou nomes de espécies necessitam também de dicionários específicos, onde seja possível consultar outros nomes científicos aceitos pela comunidade científica, ou ainda nomes populares. Com este intuito, o serviço acessa a base de dados ITIS(Integrated Taxonomic Information System)[4] [McDiarmid 1998] na busca por esse tipo de sinônimos.

5. Conclusões e Perspectivas de Longo Prazo

Este artigo apresentou a especificação e desenvolvimento de um serviço Web para ontologias, voltado para Sistemas de Biodiversidade. O trabalho é centrado no levantamento de aspectos ontológicos desses sistemas, especificando as necessidades de solicitação a um serviço de ontologias. Seu objetivo é atender a demanda por sistemas caracterizados pela multiplicidade de usuários e visões do mundo, em um contexto de uso de grandes volumes de dados multimídia na Web, contribuindo assim para questões levantadas pelo primeiro grande desafio. A validação do trabalho está sendo realizada com a ajuda de especialistas do domínio - os biólogos do Instituto de Biologia parceiros do projeto WeBios.

[4]http://www.itis.gov/

Os cinco Grandes Desafios apresentam questões que requerem pesquisa inovadora e multidisciplinar em Computação. Uma perspectiva de longo prazo requer vários tipos de consideração. Em primeiro lugar, encarados em um alto nível de abstração, se referem a cinco visões da pesquisa na área, envolvendo dados (primeiro desafio), modelos (segundo), hardware (terceiro), pessoas (quarto) e sistemas confiáveis (quinto), combinando todos os demais fatores. Sob tal ótica, uma visão de futuro envolvendo qualquer desafio pode tanto considerá-los em conjunto quanto se centrar em apenas um deles. Este artigo adota este último enfoque, concentrando-se em destacar aspectos de pesquisa que possam contribuir para resolver problemas do primeiro desafio, ao mesmo tempo que apóiam soluções para o segundo. De fato, quando se facilita o gerenciamento semântico de dados, se fornece condições para melhor testar e desenvolver os modelos que usarão tais dados.

Em segundo lugar, o primeiro desafio está intrinsecamente ligado a três fatores: (1) a proliferação de dispositivos baratos de captura de dados, (2) o aumento de software que facilite a usuários a criação e editoração de conteúdo e (3) a disseminação do uso da Web como meio preferencial de publicação de informação. Esta combinação de fatores causa o chamado fenômeno de ”dilúvio de dados”, que há anos vem inquietando não apenas a comunidade científica em Computação, mas a sociedade como um todo. O dilúvio nunca irá desaparecer – poderá, talvez, ser represado para re-aparecer sob nova forma (vide discussões a respeito no século XVI, logo após a descoberta da imprensa).

Com isto, uma terceira consideração corresponde ao sem-número de perguntas a serem feitas, cada uma das quais relativas a pontos em aberto de pesquisa em Computação. Na verdade, perspectivas de futuro podem ser analisadas a partir de perguntas que envolvam alguns desses pontos. Como, de fato, separar o joio do trigo? Quais as informações relevantes, como achá-las e filtrá-las, como mostrá-las de acordo com o contexto do usuário, o que fazer para tomar decisões a partir de tais informações, como contribuir para esclarecer os problemas associados? No caso de informações incorretas, como corrigir os dados que formaram a base para gerá-las, que dados são esses, onde se encontram, o que fazer se estão replicados? Quais (novas) métricas devem ser usadas para medir a veracidade de um resultado de uma consulta, ou de sua utilidade? De que forma garantir a integridade, a confiabilidade e a durabilidade de dados e informações? Como estabelecer limites entre a privacidade dos cidadãos que acessam os dados e as necessidades reais de proteger cidadãos contra os mais diversos tipos de crimes? Dados e conteúdo digital como um todo devem ser tratados como bens ou como recursos? Quais os fatores econômicos e sociais envolvidos em sua disponibilização?

Estas perguntas sinalizam apenas alguns dos inúmeros aspectos de pesquisa associados ao gerenciamento de grandes volumes de dados multimídia distribuídos. Cada uma delas serve de motivação para projetos de porte; propostas para solução de qualquer um desses itens causa repercussão nos demais - positiva ou negativamente. Com isto, uma quarta consideração é que sempre será preciso ter em mente os desdobramentos de soluções para o(s) desafio(s). Como é impossível que pequenos grupos tenham consciência de todos os desdobramentos, cada vez mais o tratamento dos problemas exigirá equipes de perfil intra- e inter-disciplinar. A interação entre projetos e equipes pode muito bem vir a se tornar um sexto desafio.

Uma quinta consideração concerne os enfoques atuais para enfrentar o primeiro desafio. Serviços Web se caracterizam pelo fraco acoplamento e a transparência de plata-

forma, além de facilitar a interação entre aplicações escritas em diversas linguagens e executando em diferentes plataformas, sendo atualmente a melhor solução para as questões de interoperabilidade inerentes ao desafio. Da mesma forma, ontologias são preconizadas para resolver problemas semânticos, de vocabulário e de contexto. Novos modelos de dados e linguagens de consulta, estruturas dinâmicas de armazenamento e indexação, novos algoritmos de mineração e compactação de dados, e padrões de publicação e intercâmbio de dados são outros aspectos tecnológicos e científicos associados. Não se pode esquecer, tampouco, a evolução em hardware e nas tecnologias de comunicação, que vêm facilitando o armazenamento e transmissão de dados e agilizando o seu processamento. Todas essas pesquisas são relevantes, respondem a demandas reais do desafio e contribuem para resolvê-lo. No entanto, qualquer solução mais ampla precisará envolver os usuários – a população mundial – para criar soluções que sejam efetivamente adotadas e até para atender demandas reais porém não vislumbradas atualmente. Este envolvimento requer, dentre outros, uma revolução nas técnicas atuais de análise de requisitos da Engenharia de Software e de avaliação (de desempenho, de usabilidade, de confiabilidade).

Finalmente, a sexta consideração abrange os cinco desafios e envolve educação continuada (de usuários leigos e especialistas, de cientistas das diversas áreas envolvidas, e principalmente de professores). Não há nenhuma possibilidade de resolver qualquer desafio sem promover modificações em nossos curriculos e na maneira de ensinar de forma a envolver aspectos multidisciplinares e estimular a curiosidade. Isto se aplica não apenas à Computação como um todo, mas a todas as outras disciplinas. E, ao criar material didático, precisamos ter em mente que também estamos aumentando o volume de dados multimeios distribuídos, realimentando os problemas do primeiro desafio.

Agradecimentos Agradecimentos ao apoio financeiro recebido pelo CNPq, CAPES, FAPESP (processo 05/57424-0) e à *Microsoft Research* financiadora do projeto *WeBios*.

Referências

Alani, H., Brewster, C., and Shadbolt, N. (2006). Ranking Ontologies with AKTiveRank. In *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 1–15. Springer.

Alani, H., Harris, S., and O'Neill, B. (2005). OntologyWinnowing: A Case Study on the AKT Reference Ontology. In *CIMCA/IAWTIC*, pages 710–715. IEEE Computer Society.

Antoniou, G. and van Harmelen, F. (2003). Web Ontology Language: OWL. In Staab, S. and Studer, R., editors, *Handbook on Ontologies in Information Systems*, pages 76–92.

Bruijn, J., Martin-Recuerda, F., Manov, D., and Ehrig, M. (2004). State-of-the-art survey on Ontology Merging and Aligning. Technical report, SEKT project D4.2.1.

Carroll, J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. (2004). Jena: implementing the semantic web recommendations. In *WWW Alt. '04: Proc. of the 13th international World Wide Web*, pages 74–83. ACM Press.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., and Sachs, J. (2004). Swoogle: a Search and Metadata Engine for the Semantic Web. In

CIKM '04: Proc. of the thirteenth ACM international conference on Information and knowledge management, pages 652–659, New York, NY, USA. ACM Press.

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

Duke, M. and Patel, M. (2003). An Ontology Server for Agentcities.NET. Agentcities Task Force Technical Note.

E. Jiménez, R. Berlanga, I. S. M. J. A. R. D. (2005). OntoPathView: A Simple View Definition Language for the Collaborative Development of Ontologies. *In B. López et al. (Eds.): Artificial Intelligence Research and Development*, pages 429–436.

Felicíssimo, C. H. (2004). Interoperabilidade Semântica na Web: Uma Estratégia para o Alinhamento Taxonômico de Ontologias. Master's thesis, PUC-Rio de Janeiro.

Gennari, J. H., Musen, M. A., Fergerson, R., Grosso, W. E., Crubzy, M., Eriksson, H., Noy, N. F., and Tu, S. W. (2003). The Evolution of Protege: An Environment for Knowledge-Based Systems Development. *International Journal of Human-Computer Studies*, 58(1):89–123.

Gruber, T. (1995). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928.

Hartmann, J., Sure, Y., Haase, P., Palma, R., and Suárez-Figueroa, M. C. (2005). OMV – Ontology Metadata Vocabulary. In Welty, C., editor, *ISWC 2005 - In Ontology Patterns for the Semantic Web*.

Kalfoglou, Y. and Schorlemmer, M. (2003). Ontology Mapping: the State of the Art. *Knowledge Engineering Review*, 18(1):1–31.

Kong, H., Hwang, M., and Kim, P. (2005). A New Methodology for Merging the Heterogeneous Domain Ontologies Based on the WordNet. In *NWESP '05: Proc. of the International Conference on Next Generation Web Services Practices*, page 235, Washington, DC, USA. IEEE Computer Society.

Lee, J. (2003). An Application Programming Interface for Ontology. IBM T. J. Watson Research Center. Document from SNOBASE v.1.0 release documentation.

Li, Y., Thompson, S., Tan, Z., Giles, N., and Gharib, H. (2003). *Beyond Ontology Construction; Ontology Services as Online Knowledge Sharing Communities*.

McDiarmid, R. (1998). The Integrated Taxonomic Information System. In *Proc. of the Taxonomic Authority Files Workshop*.

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.

Noy, N. F. and Musen, M. A. (2004). Specifying Ontology Views by Traversal. In McIlraith, S. A., Plexousakis, D., and van Harmelen, F., editors, *International Semantic Web Conference*, volume 3298 of *Lecture Notes in Computer Science*, pages 713–725.

Parr, C. S., Parafiynyk, A., Sachs, J., Ding, L., Dornbush, S., Finin, T. W., Wang, D., and Hollander, A. (2006). Integrating Ecoinformatics Resources on the Semantic Web. In *Proc. in 15th International Conference on World Wide Web*, pages 1073–1074. ACM.

Patel, C., Supekar, K., Lee, Y., and Park, E. K. (2003). OntoKhoj: a Semantic Web Portal for Ontology Searching, Ranking and Classification. In *WIDM '03: Proc. of the 5th ACM international workshop on Web information and data management*, pages 58–61, New York, NY, USA. ACM Press.

Prud'hommeaux, E. and Seaborne, A. (2006). SPARQL Query Language for RDF. Technical report, World Wide Web Consortium - W3C.

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

Seaborne, A. (2003). RDQL: A Query Language for RDF. Technical report, World Wide Web Consortium - W3C.

Shvaiko, P. and Euzenat, J. (2004). A Survey of Schema-based Matching Approaches. Technical Report DIT-04-087, University of Trento.

Suguri, H., Kodama, E., Miyazaki, M., Nunokawa, H., and Noguchi, S. (2001). Implementation of FIPA ontology service. In Workshop on Ontologies in Agent Systems, 5th International Conference on Autonomous Agents.

Torres, R. S., Medeiros, C. B., Gonçalves, M. A., and Fox., E. A. (2006). A Digital Library Framework for Biodiversity Information Systems. *International Journal on Digital Libraries*, 6(1):3 – 17.

Tury, M. and Bieliková, M. (2006). An Approach to Detection Ontology Changes. In *ICWE '06: Workshop proceedings of the sixth international conference on Web engineering*, page 14, New York, NY, USA. ACM Press.

Volz, R., Oberle, D., and Studer, R. (2003). Implementing Views for Light-Weight Web Ontologies. In *Proc. of Int. Database Engineering and Application Symposium (IDEAS)*, Hong Kong, China.

WeBios (2005). WeBios: Web Service Multimodal Tools for Strategic Biodiversity Research, Assessment and Monitoring. Home Page: http://www.lis.ic.unicamp.br/projects/webios.