Aondê: Um Serviço Web de Ontologias para Interoperabilidade em Sistemas de Biodiversidade

Jaudete Daltio[1], **Claudia Bauzer Medeiros (orientadora)**[1]

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

jaudete@gmail.com, cmbm@ic.unicamp.br

Abstract. *Research in biodiversity associates data on living beings and their habitats, constructing sophisticated models and correlating several kinds of heterogeneous data. Such data are provided by research groups with different vocabularies, methodologies and goals, which hampers their cooperation. Ontologies are being proposed as one of the means to solve heterogeneity problems. However, this gives birth to new challenges to manage and share ontologies. This dissertation specified and developed a new kind of Web Service, whose goal is to contribute to solve such problems. The service supports a wide range of operations on ontologies, and was implemented and validated with real case studies in biodiversity, for large ontologies. The dissertation is available on UNICAMP digital library.*

Resumo. *A pesquisa em biodiversidade associa dados sobre seres vivos e seus habitats, construindo modelos sofisticados e correlacionando vários tipos de dados heterogêneos. Tais dados são providos por grupos de pesquisa com vocabulários, metodologias e objetivos distintos, dificultando a cooperação entre eles. O uso de ontologias vem sendo proposto para solucionar problemas de heterogeneidade, o que gera novos desafios de gerenciamento e compartilhamento de ontologias. Esta dissertação especificou e desenvolveu um novo tipo de Serviço Web cujo objetivo é contribuir para a solução de tais desafios. O serviço provê um amplo espectro de operações para ontologias, e foi implementado e validado com estudos de caso reais em biodiversidade e ontologias volumosas. O texto completo da dissertação está disponível na biblioteca digital da UNICAMP* [1].

1. Introdução e Motivação

A pesquisa em biodiversidade é um campo multidisciplinar que requer a cooperação de vários tipos de pesquisadores. Os biólogos realizam diferentes tipos de atividades, incluindo coletas em campo, análises de dados sobre espécimes, seus habitats e correlações com outros seres vivos, e constroem modelos capazes de descrever essas interações. Os dados disponíveis são coletados em vários lugares do mundo, publicados em formatos distintos e especificados em inúmeros padrões. Este cenário é caracterizado por sua heterogeneidade intrínseca – não apenas de dados e modelos conceituais utilizados, como também de necessidades e perfis dos especialistas que analisam os dados.

[1]http://libdigi.unicamp.br/document/?code=vtls000415249

O volume de dados e a diversidade das espécies atuam como fatores complicadores deste cenário: as estimativas sobre o número de espécies existentes no mundo variam entre 10 e 100 milhões, das quais apenas 2 milhões são atualmente conhecidas. Há uma demanda crescente por mecanismos sofisticados de armazenamento, gerenciamento e processamento desses dados, visando sua análise integrada e seu compartilhamento. Uma dificuldade adicional é a heterogeneidade temporal – tanto os ecossistemas como os modelos ecológicos e a classificação taxonômica das espécies sofrem modificações ao longo dos anos, refletindo a evolução do conhecimento científico no mundo real.

Sistemas de Informação de Biodiversidade [Torres et al. 2006] representam soluções para alguns desses problemas, permitindo a análise de espécies e suas interações. Consultas típicas nesses sistemas combinam informações textuais sobre espécimes – descrições de observações, também chamadas de *registros de coletas* – e informações geográficas, caracterizando os ecossistemas onde os espécimes foram observados e a distribuição espacial das ocorrências. A interoperabilidade e a manipulação de dados heterogêneos são os maiores desafios desses sistemas.

O uso de ontologias tem sido apontado como solução para problemas de interoperabilidade. Ontologias são descrições de um modelo abstrato de termos relacionados entre si [Gruber 1995]. Modelam uma parte da realidade, suas entidades, relações taxonômicas, propriedades e restrições, e definem um entendimento comum de um domínio. Grupos distintos de pesquisadores podem cooperar por meio do compartilhamento e combinação das ontologias que criam, que refletem seu vocabulário e percepção de um domínio. Neste cenário, ontologias impõem um novo tipo de sobrecarga às aplicações: como definir suas ontologias de interesse e, principalmente, como manipulá-las. As propostas existentes não são suficientes para suprir as necessidades de gerenciamento em cenários distribuídos com múltiplas ontologias, considerando aplicações com interesses e terminologias distintos ou que manipulam o conhecimento em diferentes níveis de detalhe.

O objetivo desta dissertação foi contribuir para a solução deste problema, apresentando um Serviço Web para ontologias – Aondê[2]. Aondê foi projetado e implementado para suprir as necessidades do WeBios [WeBios 2007], um sistema de biodiversidade para acesso a dados heterogêneos e distribuídos na Web. O WeBios está sendo desenvolvido em uma parceria de pesquisadores dos Institutos de Computação e Biologia da UNICAMP.

O mestrado foi concluído em 24 meses, com artigos em: um congresso nacional QUALIS B (SEMISH 2007 [Daltio and Medeiros 2007]), um congresso internacional QUALIS A (ACM Symp. Applied Computing, SAC 2008 [Daltio et al. 2008]) e um periódico internacional QUALIS A (aceito, Information Systems [Daltio and Medeiros 2008]). O restante deste texto está organizado da seguinte forma: a Seção 2 discute alguns tópicos de pesquisa associados ao trabalho; a Seção 3 apresenta a especificação do serviço Aondê; a Seção 4 descreve aspectos de implementação e validação e, por fim, a Seção 5 apresenta as contribuições e extensões deste trabalho.

2. Aspectos de Pesquisa Considerados

A necessidade de interoperabilidade na Web gerou um grande número de pesquisas no processamento e gerência de ontologias. Apesar de vários resultados interessantes, há

[2]Aondê significa “coruja” (em Tupi), uma referência à linguagem de representação de ontologias OWL

ainda inúmeras questões em aberto, que abrangem desde linguagens de especificação e documentação de ontologias à mecanismos de construção, processamento de consultas, compartilhamento e integração de ontologias. A implementação de ferramentas também apresenta muitos desafios, em especial pela heterogeneidade de dados e ausência de padrões consensuais implementação.

Os *toolkits* existentes para manipulação de ontologias [Perez et al. 2002, Gennari et al. 2003] são, em sua maioria, auto-contidos e não permitem acesso externo por aplicações cliente. Para permitir que aplicações compartilhem ontologias é preciso adotar outros tipos de soluções: *frameworks* ou servidores de ontologias. O primeiro tipo possibilita que a codificação da aplicação seja diretamente associada à especificação do domínio, ou seja, à ontologia. Com isso, evoluções no domínio (ontologias) impactam diretamente em recodificação de parte da aplicação [Carroll et al. 2004, Lee 2003]. Servidores, por outro lado, permitem que o código da aplicação seja independente das especificações do domínio, entretanto suas funcionalidades se restringem ao acesso e à realização de consultas em ontologias [Li et al. 2003, Duke and Patel 2003]. Além disto, a maioria das propostas não considera a utilização de metadados, necessários para viabilizar a busca e o reuso de ontologias.

O serviço Aondê combina as vantagens de servidores e *frameworks*, sendo capaz de manipular conjuntos de ontologias de forma integrada. Além disto, garante uma clara separação entre o armazenamento/persistência de ontologias e as operações que as manipulam. Finalmente, considera o uso de metadados padrões de descrição. A escolha das operações de manipulação e suas implementações também apresentaram desafios de pesquisa, dada a divergência de enfoques e soluções propostos na literatura.

3. Especificação do Serviço de Ontologias – Aondê

O serviço Aondê combina as principais abordagens existentes na literatura para manipulação de ontologias, provendo um conjunto de operações cujos parâmetros podem ser definidos dinamicamente por aplicações cliente. Tais aplicações podem assim trocar, reusar, integrar e adotar ontologias publicadas na Web.

A Figura 1 ilustra a arquitetura proposta para o Aondê, composta por duas camadas principais: *Repositórios* e *Operações*. Esta última provê funções de manipulação das ontologias armazenadas na primeira camada. A *Camada de Repositórios* é composta por vários repositórios distribuídos de ontologias, acessados via Serviços Web. Esses repositórios podem ser de dois tipos: (i) *Repositórios Semânticos*, que armazenam ontologias e metadados e são construídos e gerenciados pelo Aondê e (ii) *Repositórios Externos de Ontologias*, gerenciados por outras instituições. O módulo de *Gerência dos Repositórios* garante a comunicação entre as duas camadas.

Aplicações cliente podem interagir com o Aondê tanto via a Camada de Operações quanto acessando diretamente os Repositórios Semânticos (permitindo, por exemplo, que especialistas do domínio possam consultar, validar ou atualizar ontologias nesses repositórios). A linguagem OWL (*Web Ontology Language*) [Antoniou and van Harmelen 2003] é utilizada na representação das ontologias e o padrão OMV (*Ontology Metadata Vocabulary*) [Hartmann et al. 2005] é adotado para a representação dos metadados nos Repositórios Semânticos. A escolha dos módulos da Camada de Operações foi baseada no estudo das necessidades de sistemas de biodiversidade

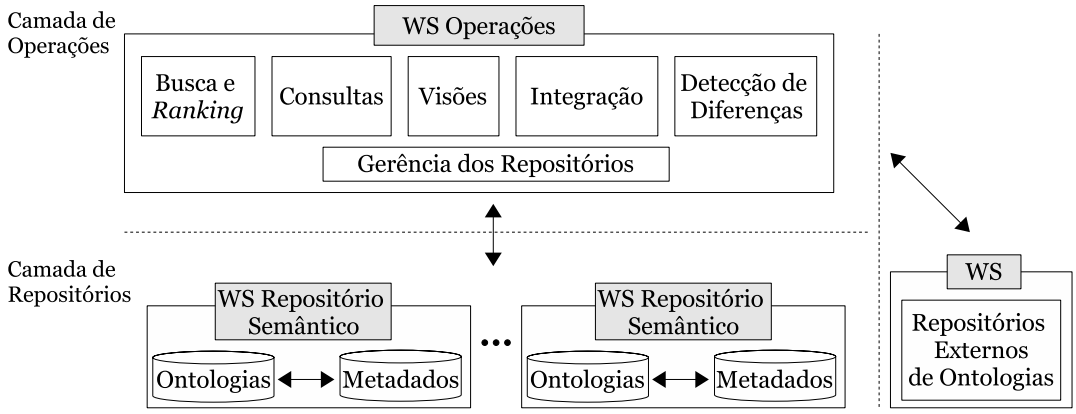

Figura 1. Arquitetura do Serviço de Ontologias

(e.g., SinBiota [3], Spire [Parr et al. 2006] e GBIF [4]) e complementada pelos requisitos identificados pelos biólogos parceiros do projeto WeBios. Os módulos que produzem novas ontologias utilizam Repositórios Semânticos de destino para armazenar o resultado de suas execuções. As seções subsequentes descrevem os principais aspectos de cada módulo do serviço.

Gerência dos Repositórios. Módulo responsável pelo gerenciamento em memória das ontologias manipuladas pelo serviço e pela mediação entre os outros módulos de operações e os Repositórios Semânticos. Permite inserção e eliminação de ontologias e metadados, atualização de ontologias (novas versões) e substituição de metadados. Cada ontologia possui um identificador único em seu repositório, composto pelo par: $<$*idOnto, URLRep*$>$, em que *idOnto* representa o identificador da ontologia no repositório endereçado por *URLRep*.

Busca e *Ranking*. Módulo que realiza a busca, em um conjunto de repositórios, por ontologias que possuam certos termos de interesse. Como resultado, retorna um conjunto de identificadores de ontologias que contém classes ou instâncias cujos nomes casam (exata ou parcialmente) com tais termos. O módulo permite dois tipos de operações: a busca por um conjunto de ontologias (com *ranking*) e a busca por uma ontologia específica (sem *ranking*). O *ranking* é baseado em um conjunto de métricas que analisam as estruturas internas de cada ontologia, sendo que cada métrica possui um peso associado. As invocações da busca possuem as formas:

$$BuscaRank(\{termo\}, \{peso\}, \{URLRep\}, URLRepDestino),$$
$$Busca(termo, \{URLRep\}, URLRepDestino),$$
$$BuscaTaxon(taxon, \{diretivas\}, \{URLRep\}, URLRepDestino),$$

em que $\{termo\}$ representa o conjunto de termos buscados nas ontologias e $\{URLRep\}$ representa o conjunto de repositórios (Semânticos ou Externos) designados para a realização da busca. $BuscaTaxon$ é uma busca em árvores taxonômicas biológicas, que difere de $Busca$ nos algoritmos de comparação de termos, pois exige conhecimento das

[3] São Paulo Biodiversity System, http://sinbiota.cria.org.br
[4] Global Biodiversity Information Facility, http://www.gbif.org

regras de nomenclatura taxonômica de espécies. O conjunto de valores em $\{peso\}$ é utilizado na combinação das métricas de *ranking* utilizadas (adaptadas de [Alani et al. 2006]).

Consultas. Módulo de execução de consultas a ontologias. As linguagens de consulta para ontologias RDQL [5] e SPARQL [6] são suportadas pelo serviço. Dada uma ontologia e uma consulta expressa em uma dessas linguagens, o módulo retorna o resultado da consulta no formato de saída requisitado: triplas RDF (ou partes dessas triplas) ou arquivos XML [7]. Uma invocação para este método possui a forma:

$$Consulta(\textit{idOnto, linguagem, strConsulta, inferencia, formato}),$$

onde *idOnto* representa o identificador da ontologia consultada, *linguagem* representa a linguagem utilizada na especificação da consulta descrita pela *string strConsulta*. O campo *formato* define o formato de saída do resultado e o campo booleano *inferencia* determina o uso ou não de mecanismos de inferência antes da execução da consulta.

Visões. Módulo responsável pela extração de visões de uma ontologia, usando a noção de "conceito central" proposta em [Noy and Musen 2004], representado por uma classe dessa ontologia. O resultado é uma sub-ontologia da ontologia original, contendo apenas a classe de interesse e os elementos relacionados a ela por meio de diretivas especificadas. Uma invocação deste módulo possui a forma:

$$Visao(\textit{idOnto, conceito,}\{\textit{diretivas}\}\textit{,URLRepDestino}),$$

em que *idOnto* representa o identificador da ontologia fonte, *conceito* refere-se ao nome da classe que representa o conceito central da visão e o conjunto $\{diretivas\}$ especifica quais elementos relacionados ao conceito central devem ser incluídos na visão (propriedades, axiomas, instâncias, por exemplo).

Integração. Módulo responsável por integrar duas ontologias. As ferramentas de integração existentes apenas procuram por mapeamentos entre elementos do mesmo tipo. Entretanto, é comum que ontologias difiram em granularidade na descrição das entidades do domínio – uma classe pode representar semanticamente o mesmo conceito que uma instância em outra ontologia. Por esta razão, além dos mapeamentos entre elementos do mesmo tipo, este módulo busca por mapeamentos entre classes e instâncias. A abordagem de integração adotada é o alinhamento, que produz uma nova ontologia preservando integralmente as ontologias originais e materializando os mapeamentos existentes entre seus elementos. Uma invocação a este módulo possui a forma:

$$Integracao(\textit{idOntoA, idOntoB, } \alpha, \beta\textit{, confiabilidadeMin, URLRepDestino}),$$

em que *idOntoA* e *idOntoB* representam os identificadores das ontologias integradas. O campo *confiabilidadeMin* corresponde à confiabilidade mínima que um mapeamento identificado deve ter para ser materializado na ontologia alinhada. A *confiabilidade* de um mapeamento é representada por um número entre 0 e 1, correspondendo à similaridade entre os elementos que participam do mapeamento. Essa similaridade é

[5] http://www.w3.org/Submission/2004/SUBM-RDQL-20040109
[6] http://www.w3.org/TR/rdf-sparql-query
[7] http://www.w3.org/TR/2006/CR-rdf-sparql-XMLres-20060406

calculada combinando técnicas de comparação de *strings* (métrica Jaro [Cohen et al. 2003]), dicionários de sinônimos (ITIS [8] e WordNet [9]) e comparações estruturais entre os elementos nas ontologias (superclasses, subclasses, propriedades e axiomas similares), utilizando os pesos α e β.

Detecção de Diferenças. Módulo que compara duas ontologias, enumerando as diferenças existentes em suas estruturas (classes e propriedades) e conteúdos (instâncias). O resultado é um arquivo XML contendo as diferenças detectadas. Uma invocação a este módulo possui a forma:

$$Diferenca(\textit{idOntoA}, \textit{idOntoB}, \alpha, \beta, \textit{confiabilidadeMin}),$$

em que *idOntoA* e *idOntoB* representam os identificadores das ontologias comparadas. O campo *confiabilidadeMin* corresponde à confiabilidade mínima que um mapeamento identificado deve ter para ser considerado uma modificação entre as ontologias comparadas. Esse cálculo é realizado de forma análoga ao Módulo de Integração.

4. Implementação e Validação

Um protótipo do serviço Aondê foi implementado na linguagem Java e utilizando o SGBD PostgreSQL para persistência dos repositórios semânticos. O acesso e a navegação ao conteúdo das ontologias é provido pelo *framework* Jena [Carroll et al. 2004]. A implementação de Serviços Web utiliza o *framework* de código aberto Apache Axis. Esse *framework* consiste em uma implementação Java de um servidor SOAP e vários utilitários e APIs para criação e publicação de Serviços Web.

As ontologias para os teste do serviço Aondê foram construídas utilizando a ferramenta Protégé [Gennari et al. 2003]. Essa ferramenta possui um *plug-in* que permite a manipulação de ontologias na linguagem OWL, com uma interface gráfica para a criação de classes, propriedades, instâncias e axiomas em lógica descritiva. O serviço foi validado em vários experimentos e estudos de caso, ilustrando o uso de cada um dos módulos do serviço Aondê na solução de consultas relevantes a dados de biodiversidade. Os exemplos são baseados em dados reais, coletados para pesquisas dos biólogos parceiros do projeto WeBios ao longo de 3 anos. O estudo está relacionado com interações entre insetos (moscas e borboletas) e inflorescências de plantas. Foram criadas 3 ontologias de coletas diferentes, armazenadas em repositórios distribuídos. A maior delas era composta por 5700 elementos, representados por cerca de 24.000 tuplas RDF.

Algumas das consultas testadas não podem ser resolvidas pela maioria das ferramentas disponíveis, cujas operações são isoladas e se restrigem a uma ou duas ontologias. Os exemplos da dissertação, por outro lado, exigem uma composição de invocações ao Aondê, combinando a criação dinâmica de ontologias intermediárias (usando visões) e a integração com ontologias fornecidas por terceiros, em portais na Web. Um exemplo de consulta não permitida em outras propostas é "Retorne as espécies de borboletas que se alimentam de plantas da espécie *Chromolaena odorata*". Aparentemente simples, esta consulta requereu combinar três ontologias diferentes, devido a questões como versões distintas de classificações taxonômicas ou ausência de informação sobre "borboleta" (um

[8] http://www.itis.gov
[9] http://wordnet.princeton.edu

termo leigo), já que os dados de coleta referenciavam *Lepidoptera* (o termo científico correspondente à classificação taxonômica das borboletas). Esta consulta envolveu duas buscas Web por ontologias externas, uma extração de visões e duas integrações sucessivas para, finalmente, permitir a execução da consulta desejada em SPARQL. Mais detalhes na dissertação e em [Daltio and Medeiros 2008].

5. Contribuições e Extensões

A dissertação especificou e implementou o Aondê - um Serviço Web para gerenciamento de ontologias, visando interoperabilidade de aplicações heterogêneas e distribuídas na Web. Embora a implementação realizada esteja direcionada para o domínio de biodiversidade, o serviço foi especificado de forma genérica, podendo ser estendido e adotado por outras aplicações que possuam necessidades de interoperabilidade similares. As principais contribuições da dissertação, teóricas e práticas, são:

- Levantamento e comparação de técnicas e ferramentas para manipulação de ontologias propostas na literatura;
- Especificação de um Serviço de Ontologias, modular e extensível, caracterizado por: (i) utilizar repositórios distribuídos para armazenar e gerenciar ontologias e seus metadados; (ii) acessar esses repositórios por meio de protocolos padrão de Serviços Web; e (iii) disponibilizar para aplicações cliente um conjunto de operações que permitem a manipulação integrada de conjuntos de ontologias. Tais operações englobam e estendem as principais propostas da literatura;
- Implementação do Serviço Web especificado – Aondê – voltado às necessidades de pesquisadores em biodiversidade e validado com estudos de caso reais. Enquanto a maioria dos trabalhos correlatos se restringe a exemplos didáticos, estes estudos de caso distinguem-se pelo tamanho das ontologias consideradas e pela complexidade da composição de operações que exigem.

Há várias extensões possíveis, tanto em pesquisa quanto em implementação, dentre as quais: a incorporação de mecanismos internos para gerenciar a procedência de uma ontologia e a gerência de versões; acoplar outros raciocinadores mais completos ao módulo de consultas; usar abordagens heurísticas no módulo de detecção de diferenças; realizar buscas por metadados no módulo de busca; implementar múltiplos conceitos centrais no módulo de visões e incorporar outras técnicas de similaridade e abordagens de integração no módulo de integração.

Agradecimentos: Apoio financeiro recebido da CAPES, FAPESP (processo 05/57424-0), e *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.

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

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.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proc. of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, pages 73–78, Mexico.

Daltio, J. and Medeiros, C. B. (2007). Um serviço de ontologias para sistemas de biodiversidade. In *XXXIV SEMISH: Seminário Integrado de Software e Hardware*, pages 2143–2157, Rio de Janeiro, Brazil.

Daltio, J. and Medeiros, C. B. (2008). Aondê: An Ontology Web Service for Interoperability across Biodiversity Applications. *Information Systems*. Accepted for publication, 29 pages, doi:10.1016/j.is.2008.02.001.

Daltio, J., Medeiros, C. B., Jr, L. C. G., and Lewinsohn, T. (2008). A framework to process complex biodiversity queries. In *Proc. ACM Symposium on Applied Computing (ACM SAC)*.

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

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 *ISWC 2005 - In Ontology Patterns for the Semantic Web*, Galway, Ireland.

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. G., Tan, Z., Giles, N., and Gharib, H. (2003). Beyond Ontology Construction; Ontology Services as Online Knowledge Sharing Communities. In *Intern.l Semantic Web Conference - ISWC*, v. 2870 of LNCS, pages 469–483. Springer.

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 Inter. Conference on World Wide Web*, pages 1073–1074. ACM.

Perez, A. G., Angele, J., Lopez, M. F., Christophides, V., Stutt, A., and Sure, Y. (2002). A survey on ontology tools. Deliverable 1.3, EU IST Project IST-2000-29243 OntoWeb.

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.

WeBios (2007). WeBios – Web Service Multimodal Tools for Strategic Biodiversity Research, Assessment and Monitoring. Home page http://www.lis.ic.unicamp.br/projects/webios.