UM FRAMEWORK PARA COMPOSIÇÃO SEMÂNTICA DE WORKFLOWS CIENTÍFICOS

LARYSSA MACHADO DA SILVA[#1]; REGINA BRAGA[#2]; FERNANDA CAMPOS[#3]

[#]Mestrado em Modelagem Computacional, Universidade Federal de Juiz de Fora
[1]Aluna do Mestrado em Modelagem Computacional – UFJF –
silva.laryssa@ig.com.br
[2]Professora do Departamento de Ciência da Computação – UFJF –
regina.braga@ufjf.edu.br
3Professora do Departamento de Ciência da Computação – UFJF –
Fernanda.campos@ufjf.edu.br

RESUMO

Um elemento importante para a pesquisa em *e-Science*, onde se inclui pesquisas em modelagem computacional, é o *workflow* científico que, em geral, é muito extenso, composto por muitas computações, e voltado para representar um processo experimental científico, geralmente de natureza colaborativa. Por ser, em geral, tão extenso, um *workflow* científico acaba se tornando de difícil definição. Uma forma de facilitar o processo que leva à representação de um *workflow* científico é utilizando ferramentas que fazem uso da semântica para facilitar sua composição. Neste contexto, este trabalho apresenta uma proposta que tem como principal objetivo facilitar a composição de

workflows científicos, considerando projetos em modelagem computacional, realizando buscas semânticas de serviços Web semânticos e incorporando estes na definição dos *workflows*.

PALAVRAS-CHAVE: *workflow* científico, serviço Web semântico, *e-Science*, composição.

A FRAMEWORK FOR SEMANTIC COMPOSITION OF SCIENTIFIC WORKFLOWS

ABSTRACT

An important element in research on e-Science, which includes research in computational modeling, is the scientific workflow that, in general, is very extensive, comprising many computations, and forward to represent an experimental scientific procedure, generally collaborative. Because it is generally so extensive, a scientific workflow eventually becomes difficult to define. One way to facilitate the process that leads to the representation of a scientific workflow is using tools that make use of semantics to facilitate its composition. In this context, this paper presents a proposal whose main objective is to facilitate the composition of scientific workflows, considering projects in computational modeling, performing semantic search of semantic web services and incorporating these into the definition of the workflows.

KEY WORDS: scientific workflow, semantic Web service, e-Science, composition

1 - INTRODUÇÃO

A Ciência da Computação revolucionou a pesquisa científica e hoje é reconhecida como o terceiro pilar a sustentá-la, juntamente com os pilares da teoria e da experimentação [9]. A utilização de recursos computacionais no desenvolvimento da pesquisa beneficia o trabalho das comunidades científicas facilitando o compartilhamento de dados e serviços computacionais, além de contribuir para a construção de uma infraestrutura de dados e de uma comunidade científica distribuída [6]. Este contexto em que a computação se torna parte integrante e imprescindível para o sucesso na realização de pesquisas científicas das mais variadas áreas, é o contexto de *e-Science* [4], em que a ciência é realizada com o apoio computacional, se tornando assim, mais eficiente.

As atividades de *e-Science* estão crescendo por todo o mundo, acompanhadas por uma proliferação de dados e ferramentas. Isto traz novos desafios, por exemplo, como entender e organizar esses recursos, como compartilhar e reusar experimentos bem sucedidos (ferramentas e dados), e como prover interoperabilidade entre dados e ferramentas de diferentes locais e utilizados por usuários com perfis distintos [2].

No contexto de *e-Science*, as tecnologias de *workflow* fornecem ambientes para a resolução de problemas de natureza científica, facilitando a criação e execução de experimentos a partir do uso de uma grande quantidade de dados e serviços disponíveis [15]. Os *workflows* científicos são mecanismos para representar computacionalmente os experimentos científicos. *Workflows* científicos estão sendo cada vez mais adotados como meios para especificar e coordenar a execução de experimentos multidisciplinares. Eles permitem a representação e execução de tarefas que usam dados e ferramentas heterogêneos.

Neste contexto, a utilização de ferramentas semânticas na busca de recursos para compor *workflows* científicos é muito útil, pois nem sempre é trivial compor um processo escolhendo com exatidão os recursos para cada tarefa que o compõe, principalmente no cenário atual em que esta composição normalmente é realizada escolhendo funcionalidades entre um grande número de recursos disponíveis na Web.

O objetivo deste artigo é, portanto, discutir a utilização de recursos da Web Semântica como forma de facilitar a descoberta e composição de aplicações e ferramentas para o desenvolvimento de *workflows* científicos. Nossa contribuição consiste no desenvolvimento de um *framework*, que utiliza recursos semânticos, como ontologias e serviços Web semânticos, para auxiliar na composição de *workflows* científicos, tornando assim, este processo mais simples e eficaz. Este trabalho está organizado da seguinte forma: é apresentada a fundamentação teórica; algumas propostas e trabalhos relacionados serão analisados; a proposta de desenvolvimento de um *framework* que agregue a semântica no desenvolvimento de *workflows* científicos é formulada e sua arquitetura é apresentada; finalmente, conclusões e trabalhos futuros são apresentados.

2 - FUNDAMENTAÇÃO TEÓRICA

O termo *e-Science* foi introduzido, no Reino Unido, pelo então diretor geral dos conselhos de pesquisa, *Sir* John Taylor, para encapsular as tecnologias necessárias ao suporte à pesquisa colaborativa e multidisciplinar que emergiu em vários campos da ciência [4]. Este termo é geralmente empregado para descrever o desenvolvimento de infraestruturas de serviços de software capazes de prover acesso a facilidades remotas, recursos computacionais remotos, armazenamento de informações em bancos de dados dedicados e disseminação e compartilhamento de dados, resultados e conhecimento. Em essência, “um ambiente de *e-Science* deve permitir o compartilhamento de recursos coordenados em larga-escala entre comunidades dinâmicas de indivíduos, grupos, laboratórios e instituições, permitindo

o gerenciamento cooperativo de facilidades (equipamentos, instrumentação, experimentos etc.) e análise colaborativa de produtos (dados) oriundos dessas facilidades" [12]. São características de *e-Science:* o acesso a uma vasta coleção de dados, utilização de recursos computacionais em larga escala, utilização de recursos heterogêneos e dinâmicos de múltiplas organizações, e *workflows* dinâmicos [9].

O conceito de *workflow* pode ser definido, em linhas gerais, como um modelo que define o fluxo de processos ou fluxo de tarefas coordenadas e encadeadas usando um plano sistemático. Para controlar e facilitar a utilização dos diversos recursos disponíveis no contexto de *e-Science* e possibilitar a análise eficiente de grande volume de dados são definidos *workflows* com características específicas para *e-Science*, que são os chamados *workflows* científicos. *Workflows* científicos geralmente são muito extensos, compostos por muitas computações, e voltados para explicitar representações de processos experimentais científicos, geralmente de natureza colaborativa [3].

Tradicionalmente, a Web vem sendo vista como uma fonte de informações distribuída. O surgimento da tecnologia de serviços Web permitiu que a Web se tornasse também uma fonte distribuída de funcionalidades [16]. Os serviços Web possibilitam a comunicação entre aplicações e a interoperabilidade por meio do uso de padrões Web [1]. Uma definição comumente usada de serviço Web é a seguinte: "Um serviço Web é um sistema de software desenvolvido para dar suporte à interação entre máquinas por meio de uma rede. Ele apresenta uma interface descrita em um formato processável por máquina (especificamente WSDL). Outros sistemas interagem com o serviço Web da forma definida em sua descrição, usando mensagens SOAP [19]".

Considerando o crescimento da Web tanto em tamanho, como em diversidade, há uma necessidade crescente de automatizar os aspectos relacionados aos serviços Web, como sua descoberta, execução, seleção e composição automáticas. O uso de tecnologias da Web Semântica como serviços Web semânticos e ontologias, pode auxiliar na construção de ferramentas que permitam este processo de automatização.

Dentre as técnicas e fundamentos necessários para concretizar os objetivos da Web semântica, as ontologias têm um papel fundamental. Uma definição bastante utilizada para definir ontologia é a de que "uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada" [18]. Uma linguagem para a

definição de ontologias é a OWL (*Web Ontology Language*) [13], que é a linguagem de representação de ontologias padrão da Web semântica atualmente, sendo recomendada pelo W3C.

A aplicação conjunta da Web semântica e do serviço Web, a fim de gerar serviços Web mais inteligentes, deu origem a serviços Web semânticos [10]. Um serviço Web semântico consiste em um serviço Web com uma descrição semântica que permita que a descoberta, composição e invocação do serviço possam ser feitas de forma automática, por exemplo, com o uso de agentes inteligentes capazes de processar a informação semântica disponibilizada [16].

3 - TRABALHOS RELACIONADOS

A criação de *workflows* científicos não é uma tarefa trivial. É difícil definir formalmente experimentos científicos. A tarefa de criar *workflows* científicos com serviços Web disponíveis se torna ainda mais complexa, pois os cientistas nem sempre são da área de Ciência da Computação, e por isso, não têm facilidade para utilizar recursos computacionais para definir seus experimentos científicos. Há ainda o problema de que o número de serviços disponíveis está cada vez maior e esta grande quantidade de serviços disponíveis aumenta a dificuldade para que os cientistas encontrem aqueles que realmente os interessam. Neste cenário, diversos esforços vêm sendo feitos no sentido de automatizar este processo de descoberta de serviços.

O MathWS [8] é um *broker* baseado em ontologias, para registro, descoberta e execução de serviços Web matemáticos. O protótipo do *broker* faz uso de OWL-S [14], MathML [7] e mecanismos de inferência para descoberta de serviços. Neste trabalho, o registro da descrição semântica do serviço é feito através do armazenamento, em banco de dados, dos conceitos inferidos na ontologia. A descoberta usa uma máquina de inferência para recuperar serviços que combinem com os conceitos fornecidos pelo cliente. A execução utiliza a descrição contida na

ontologia para acessar o serviço, bem como realiza a conversão dos parâmetros de conceitos semânticos para os tipos de dados esperados pelo serviço.

A proposta apresentada neste trabalho também engloba a descoberta de serviços Web por meio do uso de ontologias, fazendo uso de OWL-S e mecanismos de inferência que auxiliem na descoberta de serviços. Mas, a descoberta dos serviços se torna parte de um processo maior que envolve a composição de *workflows* científicos utilizando os serviços descobertos, e sua posterior execução.

O trabalho de [16] tem como objetivo principal a obtenção de um sistema que, a partir do emprego de agentes inteligentes e do uso intensivo de ontologias, melhore o rendimento no consumo de serviços Web e incremente a automatização das tarefas de gestão dos mesmos.

O sistema apresenta independência do domínio e da aplicação, atendendo assim a diversas aplicações em domínios distintos. Uma das tarefas fundamentais da plataforma é a gestão e manejo dos serviços Web semânticos realizando um conjunto de tarefas básicas: descoberta, seleção, composição e invocação de serviços. O funcionamento do sistema se dá completamente por meio da ação de agentes, sendo constituído por sete agentes, quatro bases de conhecimento que contêm as ontologias necessárias ao funcionamento da plataforma e três interfaces para relacionamento com entidades externas: consumidores de serviços, provedores de serviços, e desenvolvedores de software.

Este trabalho busca apoiar as tarefas de descoberta, seleção, composição e invocação de serviços por meio do uso de agentes e tecnologias semânticas. Desta forma, difere de nossa proposta no sentido de que pretendemos ir além destas funcionalidades oferecidas. Nosso objetivo é permitir a realização destas mesmas tarefas oferecidas pelo sistema proposto por [16] e também permitir a composição de *workflows* científicos com estes serviços, porém nossa proposta não faz uso de agentes para a realização das funcionalidades propostas.

4 - ARQUITETURA PROPOSTA

As tecnologias de *workflows* científicos têm contribuído com o trabalho dos pesquisadores, facilitando a criação e execução de experimentos que fazem uso de diversos recursos. No contexto de *e-Science*, podem ser considerados como principais benefícios da utilização de *workflows*, maior eficiência e confiabilidade nos processos experimentais, compartilhamento de informações, cooperação entre diversos grupos de pesquisa e possibilidade de repetição e reprodução de procedimentos.

A utilização de serviços Web semânticos pode ser muito útil na tarefa de combinar serviços diversos para compor um *workflow* científico. Os serviços Web semânticos permitem que sua descoberta, seleção, composição e invocação sejam realizadas de forma automática (sem intervenção humana) ou semiautomática (com pouca intervenção humana) por meio do uso de ferramentas computacionais capazes de interpretar semanticamente estas descrições [11]. Com o uso de inferências semânticas baseadas em ontologias e das descrições semânticas destes serviços é possível encontrar aqueles "semanticamente compatíveis" que podem ser conectados à entrada ou saída de um outro serviço, durante a composição de um *workflow*. Com o objetivo de auxiliar na descoberta de serviços Web para a composição de *workflows* científicos, propomos o desenvolvimento de um *framework* que realiza busca semântica de serviços e compõe estes serviços definindo assim, um *workflow* científico.

Na Figura 1 é representada a arquitetura em camadas do *framework* proposto, o que permite um melhor entendimento das etapas realizadas pelo mesmo para alcançar seu objetivo principal que é o de descobrir serviços Web semânticos e compor um *workflow* científico a partir destes. Podemos dividir o *framework* em dois blocos: o bloco do cliente, que invoca o serviço, e o bloco que contém as camadas que compõem o *framework*.

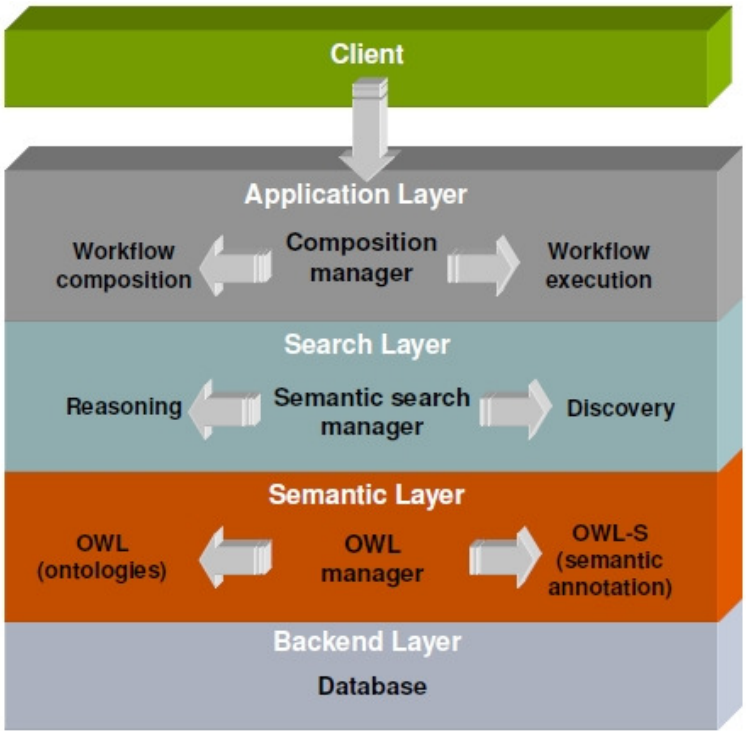

FIGURA 1 - Arquitetura em camadas do *framework* proposto

O bloco do cliente implementa uma interface por meio da qual o usuário interage com o *framework*, explicitando as necessidades e objetivos relacionados ao *workflow* que ele pretende criar. Esta interface consiste em um formulário por meio do qual o usuário especifica a descrição semântica de cada tarefa que compõe o *workflow* e também sua ordem de execução.

As demais camadas do *framework* realizam a busca e composição dos serviços científicos. O *framework* é composto por quatro camadas, sendo elas:

- **Backend Layer** é a camada que contém a base de dados relacional onde são encapsuladas (armazenadas) as ontologias de domínio (OWL) e as descrições semânticas de serviços Web (OWL-S). Esta camada é responsável por controlar o acesso e gerenciar as conexões das demais camadas à base de dados. O banco de dados relacional utilizado para o armazenamento das ontologias é o DB2 [5].
- **Semantic Layer** é a camada onde serão gerenciados os acessos das demais camadas às ontologias (OWL) e descrições semânticas de serviços (OWL-S) (anotações semânticas dos serviços). Além disso, esta camada é responsável pelo encapsulamento das ontologias em bancos de dados

relacionais, o que permite que sejam realizadas consultas e inferências sobre elas. Atualmente utilizamos o *framework* SOR [5] para auxiliar no encapsulamento das ontologias em bancos de dados relacionais e auxiliar também na realização de consultas SPARQL [17] e inferências semânticas nas ontologias de domínio. Rotinas de consultas e inferências sobre ontologias e descrições semânticas são implementadas por esta camada.

- **Search Layer** é a camada que executa a busca semântica e a descoberta de serviços de acordo com as especificações do usuário. Esta camada solicita a realização de inferências (*reasoning*) na ontologia de domínio à *Semantic Layer*, enviando como parâmetros as especificações semânticas definidas pelo usuário. São obtidos, como resultados desta inferência, termos relacionados à especificação inicial, ampliando assim, as possibilidades de busca de serviços. As informações fornecidas pelo usuário e as obtidas por meio de inferências são utilizadas para a realização de pesquisa semântica em repositórios distribuídos para encontrar serviços Web compatíveis semanticamente com cada tarefa do *workflow* (*discovery*). Para encontrar os serviços, as descrições semânticas (OWL-S) dos serviços disponíveis nos repositórios são analisadas e comparadas com os dados semânticos fornecidos pelo usuário. Como resultados da busca realizada nos repositórios podem ser encontrados vários serviços que atendam às especificações de uma dada tarefa.
- **Application Layer** é a camada responsável pela composição do *workflow* científico utilizando os serviços descobertos por meio das ferramentas semânticas. Os serviços descobertos têm sua compatibilidade estrutural analisada e possíveis composições desses serviços, que representem o *workflow* especificado inicialmente, são disponibilizadas como opções para o usuário, que por sua vez, escolhe a que mais lhe interessar e solicita a geração do *workflow* a partir dela. Os serviços são encapsulados em componentes de execução de um *workflow*. Estes componentes são incorporados no *workflow* gerando assim uma versão executável a partir da composição escolhida, que é então disponibilizada para ser executada pelo cliente.

Para a composição de *workflows* científicos com os serviços descobertos, estamos utilizando inicialmente para testes, atores do *framework* Kepler [6], onde os

serviços estão sendo encapsulados, permitindo a composição do *workflow* executável a partir da organização destes atores. Desta forma, a execução do *workflow* poderá ser realizada utilizando-se o próprio Kepler [6]. No entanto, a composição semântica é independente do sistema de gerenciamento de *workflow* científico utilizado.

O estágio atual de desenvolvimento da arquitetura proposta conta com um protótipo do *framework*, desenvolvido como um serviço Web. O protótipo é capaz de receber do usuário, por meio de uma interface Web, as descrições semânticas relacionadas ao *workflow* e realizar inferências semânticas em uma ontologia de domínio a fim de obter termos relacionados ao fornecido pelo usuário do sistema. Analisando as anotações semânticas de serviços Web semânticos que estão armazenadas em um repositório, estes termos resultantes da inferência são utilizados para encontrar (descobrir) os serviços que melhor se enquadram no contexto da descrição fornecida pelo usuário.

5 - CONCLUSÕES E TRABALHOS FUTUROS

No contexto de *e-Science*, os recursos computacionais estão se tornando cada vez mais importantes para a realização de pesquisas científicas, acompanhadas por uma proliferação de dados e ferramentas. A proliferação de recursos gera problemas e dificuldades na gestão dos mesmos e, para as atividades de *e-Science* muitos esforços estão sendo feitos no sentido de facilitar a gestão destes recursos e melhorar o desempenho das pesquisas científicas. A quantidade de serviços disponíveis na Web vem crescendo muito, portanto, a busca por ferramentas que auxiliem na descoberta, seleção, composição e invocação destes serviços também vem se ampliando.

Um componente muito importante para a pesquisa científica no contexto de *e-Science* é o *workflow* científico, que representa experimentos científicos e pode ser criado utilizando como componentes serviços Web. Neste artigo foi apresentada a proposta de um *framework* que auxilia na descoberta e composição de serviços Web

semânticos, de modo que estes sejam utilizados na modelagem de um *workflow* científico, automatizando este processo por meio do uso de tecnologias da Web semântica como ontologias e serviços Web semânticos. Estudos de viabilidade estão sendo conduzidos com o objetivo de se verificar a real aplicabilidade do *framework*. Estes estudos estão sendo realizados no domínio biológico, de geoprocessamento e modelagem computacional como um todo.

Dentro de todo este contexto em que *e-Science*, *workflows* científicos e serviços Web são partes tão importantes, o uso de tecnologias da Web Semântica, como serviços Web semânticos e ontologias, facilita a descoberta de serviços, a composição dos *workflows* e todo o processo de experimentação e obtenção de resultados.

Podemos destacar como trabalhos futuros o estudo de questões relacionadas à proveniência de dados e uso de semântica, além de técnicas de mapeamento entre ontologias que melhorem a busca por serviços Web semânticos relacionados a um dado domínio.

REFERÊNCIAS

[1] Baraka, R., Schreiner, W.: Semantic Querying of Mathematical Web Service Descriptions. Johannes Kepler University, Linz, Austria. 2006. Disponível em http://www.springerlink.com/content/c10u751k15832468/. Acesso em 10 abril 2010.

[2] Digiampietri, L. A.: Gerenciamento de *workflows* científicos em bioinformática. Tese de Doutorado em Ciência da Computação UNICAMP. 2007.

[3] Fernandes, A., Novais, R.: Sistemas de Gerência de *Workflow* em *e-Science.* 2007. Disponível em http://www.inf.puc-rio.br/~casanova/. Acesso em 13 abril 2010.

[4] Hine, C. M. New infrastructures for knowledge production: understanding E-science. 1.ed. Information Science Publishing, 306p. 2006.

[5] IBM. 2010. Disponível em http://www.ibm.com/us/em. Acesso em 20 abril 2010.

[6] Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J., Zhao, Y.: Scientific *Workflow* Management and the KEPLER System. 2004. Disponível em http://www.sdsc.edu/~ludaesch/Paper/kepler-swf.pdf. Acesso em 18 abril 2010.

[7] MathML. Mathematical Markup Language. 2008. Disponível em http://www.w3.org/TR/MathML3/. Acesso em 15 abril 2010.

[8] Matos, E. E., Nocelli, C., Braga, R., Campos F.: MathWS: Broker de Serviços Web para e-Science. 2007. Disponível em http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=7301. Acesso em 19 abril 2010.

[9] Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Murta L.: Gerenciando Experimentos Científicos em Larga Escala. In: SBC-SEMISH´08, Belém, Brasil. p.121-135. 2008.

[10] McIlraith, S., Son, T. C., Zeng, H.: Semantic web services. IEEE Intelligent Systems, 16(2), 46–53. http://www.ksl.stanford.edu/people/sam/ieee01.pdf (2001)

[11] Moran, M., Mocan, A.: WSMX – An Architecture for Semantic Web Service Discovery, Mediation and Invocation. 2004. Disponível em http://iswc2004.semanticweb.org/PID-SUWKWXSK-1090263770.pdf. Acesso em 16 abril 2010.

[12] NESC, National e-science centre. 2005. Disponível em http://www.nesc.ac.uk/. Acesso em 20 abril 2010.

[13] OWL, Ontology Web Language. 2004. Disponível em http://www.w3.org/TR/owl-guide/. Acesso em 20 abril 2010.

[14] OWL-S. 2004. Disponível em http://www.w3.org/Submission/OWL-S/. Acesso em 22 abril 2010.

[15] Pignotti, E., Edwards, P., Preece, A., Gotts, N., Polhill, G.: Enhancing workflow with a semantic description of scientific intent. 5th European Semantic Web Conference, Tenerife, Spain. 2008.

[16] Sánchez, F. G., García, R. V., Béjar, R. M., Breis, J. T. F.: An ontology, intelligent agent-based framework for the provision of semantic web services. 2009. Disponível em http://www.sciencedirect.com. Acesso em 23 abril 2010.

[17] SPARQL. 2008. Disponível em http://www.w3.org/TR/rdf-sparql-query/. Acesso em 23 abril 2010.

[18] Studer, R., Benjamins, R., Fensel, D.: Knowledge engineering: Principles and methods. Data and Knowledge Engineering. 1998. Disponível em http://www.das.ufsc.br/~gb/pg-ia/KnowledgeEngineering-PrinciplesAndMethods.pdf. Acesso em 23 abril 2010.

[19] W3C. 2004. Disponível em http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/. Acesso em 23 abril 2010.