

Ferramenta CORIDORA *Web Mapping* para Mapeamento de Esquemas em Bancos de Dados Heterogêneos

Fernando Busanello Meneghetti¹, Fabiano Gama Paes¹, Gustavo Zanini Kantorski¹

Curso de Sistemas de Informação – Universidade Luterana do Brasil – Campus Santa Maria (ULBRA) - 97020-001 – Santa Maria – RS - Brasil
{fmeneghetti, fgpaes, gustavoz} cpd.ufsm.br

Abstract. *The integrated identification to distributed heterogeneous information is a major concern in large corporations and governmental institutions. Conceptual schema mapping supports integrated access to heterogeneous databases without the requirement of database integration. The use of available Internet tools turns out as an affordable, simple, and also platform and DBMS independent, alternative for schema mapping of heterogeneous databases. This paper presents a open tool for web schema mapping in heterogeneous data sources through the Internet.*

Key words: schema mapping, heterogeneous databases, Internet, XML

Resumo. *Nos dias atuais, a identificação de informações localizadas em bancos de dados heterogêneos é decorrente da evolução tecnológica, da descentralização que vem acontecendo nas corporações e da necessidade econômica de aproveitar a grande quantidade de informações existentes, legadas de gerações anteriores. Um paradigma para acesso integrado a bancos de dados heterogêneos é o mapeamento das diferentes representações de uma entidade no mundo real. Este artigo descreve uma ferramenta web, aberta e de baixo custo, para mapeamento de esquemas em fontes de dados heterogêneas, identificando as equivalências entre os objetos das diferentes fontes.*

Palavras-chaves: schema mapping, heterogeneous databases, Internet, XML

1. Introdução

Atualmente, dentro das instituições existem fontes de dados setoriais funcionando de forma independente para atender requisitos específicos, mas com informações comuns e, de certa maneira, institucionais. Este fato é causado porque as instituições, quando projetaram seus bancos de dados, não analisaram ou anteciparam os inter-relacionamentos que poderiam ocorrer. Com o passar do tempo, as informações foram armazenadas conforme as necessidades da própria instituição e de seus usuários. Os sistemas de integração de dados têm como objetivo fornecer aos usuários uma interface uniforme para as diversas fontes de dados, autônomas, distribuídas e heterogêneas [ÖZSU, 2001]. Fontes de dados, normalmente, são estruturadas, como bancos de dados e arquivos de dados, ou semi-estruturadas como páginas web.

Neste contexto, o projeto Coridora tem por objetivo o mapeamento de esquemas de diversas fontes heterogêneas, estruturadas ou semi-estruturadas, com o objetivo de implementar o tratamento de inconsistências derivadas da representação de equivalências de um mesmo objeto do mundo real, através de mapeamento de esquemas conceituais, identificando, consistindo e comparando divergências entre os objetos equivalentes, sem prejudicar a autonomia local dos bancos de dados.

O presente artigo aborda somente de fontes de dados estruturadas, sendo que o tratamento de fontes semi-estruturadas não será abordado neste momento, tratando-se de trabalhos futuros do projeto Coridora. Este artigo apresenta uma ferramenta *web*, aberta, para mapeamento de equivalências semânticas em bancos de dados heterogêneos.

A metodologia utilizada para desenvolver a ferramenta é apresentada na seção 2. Considerações sobre a ferramenta desenvolvida e tecnologias utilizadas são apresentadas na seção 3. Trabalhos futuros e considerações finais são descritos na seção 4.

2. Integração de Esquemas

Um paradigma para acesso integrado a bancos de dados heterogêneos é o mapeamento das diferentes representações de uma entidade no mundo real, possibilitando que os usuários tenham acesso não apenas ao objeto, mas também que recebam as informações disponíveis na forma mais adequada ao seu ambiente [RIB 1995].

O mapeamento de esquemas conceituais permite que os esquemas locais sejam mantidos na sua forma original e que a autonomia local permaneça inalterada. A existência de um registro central (RC), com uma função análoga ao Esquema Conceitual Global tem como função a manutenção de cópias dos esquemas conceituais de exportação participantes e dos objetos correspondentes, disponíveis à comunidade. Contudo, sua utilização é mais ampla, uma vez que engloba funções de gerenciamento e controle da validade dos mapeamentos efetuados [KAN 2000]. Uma representação gráfica da metodologia é apresentada na figura 1.

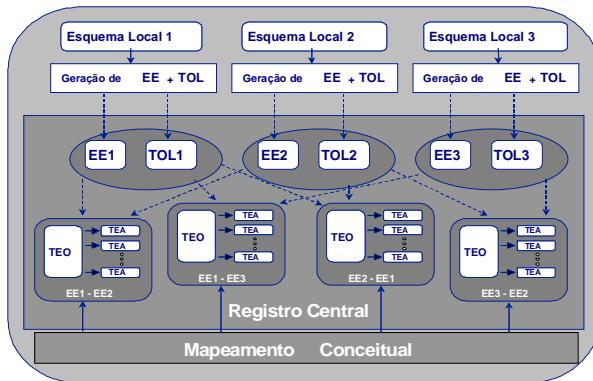


Figura 1. Mapeamento de Esquemas Conceituais.

O Esquema de Exportação (EE) contém a representação dos dados locais a serem integrados. O RC é o conjunto de todos os EEs dos vários bancos de dados exportados além de estruturas adicionais que serão descritas no decorrer deste artigo. O mapeamento propriamente dito só torna-se possível após o ingresso de pelo menos dois bancos de dados na federação [RIB 1995]. A inclusão de um membro começa pelo fornecimento do conjunto de informações que deseja disponibilizar à comunidade, através de um Esquema de Exportação (EE) e de uma Tabela de Objetos Locais (TOL). O EE é gerado a partir do esquema conceitual local e pode ser utilizado para ocultar objetos, ou atributos de objetos, devido ao caráter sigiloso da informação ou a restrições de acesso às informações do objeto. Além disso, contém informações semânticas adicionais referentes ao conteúdo das informações disponibilizadas. A TOL, contendo informações referentes ao conjunto de objetos modelados no EE correspondente, possibilita uma melhor visualização desses objetos, além de também possibilitar a inclusão de informações semânticas adicionais, acrescidas pelo projetista. Essas informações adicionais são características relativas ao objeto propriamente dito, como por exemplo, o seu nome.

A equivalência entre os objetos modelados nos esquemas conceituais é registrada através de uma Tabela de Equivalência de Objetos (TEO) correspondendo a um e a cada par de esquemas conceituais mapeados. A TEO vai sendo preenchida no decorrer do processo de comparação de esquemas. Após o encerramento da etapa de identificação de equivalências e conflitos, inicia-se o processo de mapeamento de representações equivalentes, tomando por base os objetos cujo relacionamento está identificado na TEO. Cada par de esquemas conceituais está relacionado por uma TEO local e independente, onde o relacionamento entre os objetos equivalentes é especificado.

Relacionando cada par de objetos cuja equivalência é apontada na TEO correspondente e complementando as informações necessárias é criada a Tabela de Equivalência de Atributos (TEA). Informações correspondentes a atributos equivalentes, atributos identificadores para fins de mapeamento, bem como instruções de conversão para a linguagem local ou registro de divergências a serem informadas ao usuário são armazenadas na TEA. Todo o processo de identificação de objetos equivalentes nos bancos de dados instanciados, bem como de mapeamento de atributos correspondentes e conversão de domínio para o utilizado localmente é sustentado pelas informações fornecidas pelas TEAs correspondentes aos objetos sendo acessados.

3. A Ferramenta CORIDORA *Mapping*

A ferramenta desenvolvida foi projetada usando a tecnologia de três camadas. Nessa arquitetura divide-se o ambiente em camadas logicamente independentes. O desenvolvimento em três camadas tem como objetivo retirar as regras de negócio do cliente e centralizá-las no servidor web. Todo acesso ao banco de dados ocorre de acordo com as regras implementadas no servidor web, deixando assim o cliente isolado do acesso a dados.

3.1 Camada de Persistência

A camada de banco de dados armazena todas as informações necessárias para controle do mapeamento entre os bancos de dados participantes da federação. Todas as estruturas utilizadas pela ferramenta são armazenadas em um banco de dados específico, denominado Registro Central (RC) (Figura 1). O RC é um dicionário dos dados disponíveis e das equivalências entre objetos e representações dos diversos bancos de dados a serem acessados. O modelo de dados original proposto por [KAN 2000] foi estendido para suportar novas características como o tratamento de equivalências semânticas dentro do mesmo banco de dados e suporte ao tratamento de *drivers* jdbc para diferentes bancos de dados. O modelo estendido é apresentado na figura 2.

O mapeamento entre os esquemas conceituais é efetuado a partir dos EEs e das TOLs armazenados no RC. Após a efetivação do processo de mapeamento, o usuário local passa a ter acesso às informações localizadas nos bancos de dados externos.

Como extensões do modelo original surgiram duas novas estruturas. O objeto SGBD, que foi criado com a função de dar manutenção para o cadastro de novos *driver's JDBC* que são usados para efetuar a conexão com as bases de dados e o objeto DRIVER, que foi criado com o objetivo de armazenar os dados dos novos *drivers JDBC* que poderão ser carregados pelo ambiente para que possa ser feita a conexão com a Base de dados. Pode ser usada também para carregar outros arquivos como os de licenças que alguns bancos necessitam para que juntamente com o arquivo *driver* efetue a conexão.

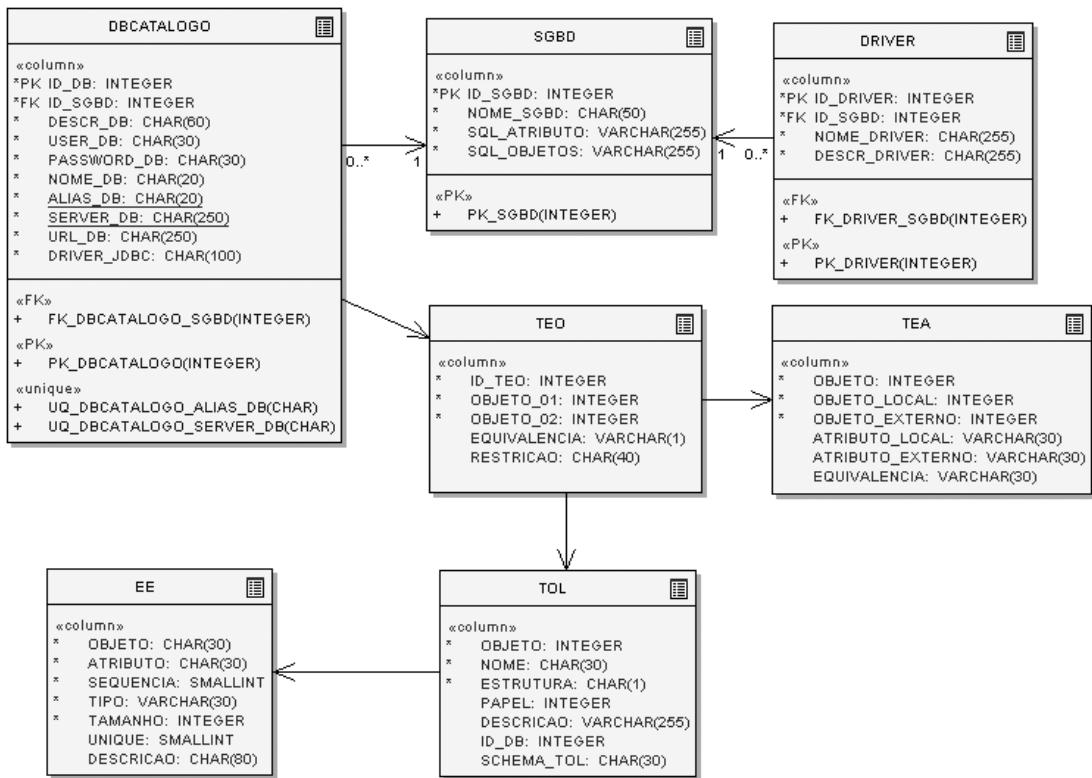


Figura 2. Modelo Estendido do Registro Central.

3.2 Camada de Aplicação

Nesta camada encontra-se o servidor *web* onde estão implementadas todas as regras que a ferramenta utiliza, desde validações de formulários, regras de conexões com o banco, inserção de dados na camada de dados, entre outras. A ferramenta fica instalada nesse servidor para disponibilizar o acesso ao cliente através do navegador. A implementação desta camada é responsável por prover uma representação das informações persistentes a outras partes do sistema. Além disso, qualquer modificação solicitada por outra parte do sistema, é realizada nesta camada.

Para a implementação foi utilizada a tecnologia JSP(Java Server Pages) uma vez que simplifica o fornecimento de conteúdo web dinâmico, reutilizando componentes pré-definidos e interagindo com componentes que utilizam *scripts* do lado do servidor, sendo possível utilizar componentes de software especiais chamados de *JavaBeans* e bibliotecas personalizadas que encapsulam funcionalidades dinâmicas e complexas. Um *JavaBean* é um componente reutilizável o qual permite ocultar o código complexo para acesso a banco de dados e outros serviços úteis para páginas da web dinâmicas [DEI 2005].

Quando uma página JSP é requisitada pelo cliente através do navegador, esta é executada pelo servidor, transformando em um *servlet*, onde o controlador interpreta a requisição do usuário, originada do formulário web acionando a execução das regras de negócio correspondente, possuindo a vantagem de poder manipular classes e interfaces na forma de pacotes Java, através do componente *JavaBeans*, com o intuito de isolar e encapsular um conjunto de funcionalidades. Por sua vez o controlador examina o resultado obtido e responde com a indicação ao formulário web que deve ser apresentado ao usuário. Para cada possível resultado da execução da regra de negócio, é selecionado um determinado formulário web, o qual é enviado ao cliente em formato de página HTML. A figura 3 ilustra esse processo.

O servidor web utilizado para armazenar a ferramenta é o *Apache Tomcat* (<http://tomcat.apache.org/>), que é um servidor de aplicações J2EE. É robusto e eficiente para ser utilizado principalmente em ambientes de produção.

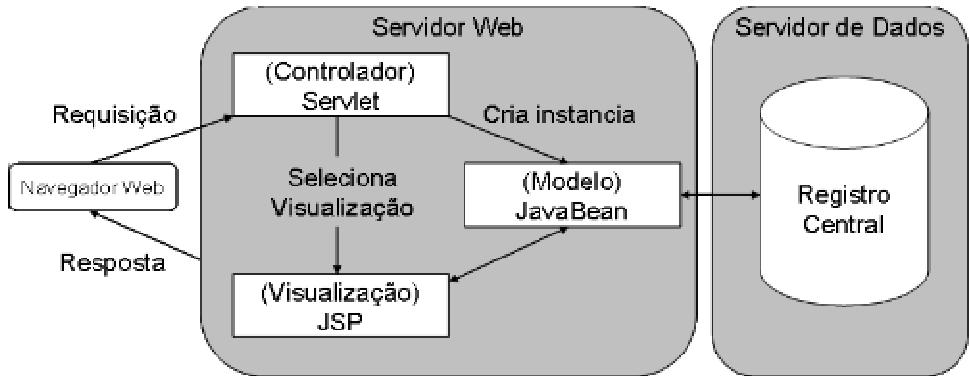


Figura 3. Implementação do Servidor Web.

3.3 Camada de Interface de Usuário

Para o desenvolvimento da aplicação foi utilizado a IDE *NetBeans* (<http://www.netbeans.org/>). O *NetBeans* é um ambiente de desenvolvimento integrado, gratuito e de código aberto para desenvolvedores de software. Oferece as ferramentas necessárias para criar aplicativos profissionais de desktop, empresariais, web, móveis e multiplataformas.

Para realizar o mapeamento o usuário deve-se logar no ambiente Coridora. Após o *logon* o usuário está habilitado para adicionar uma fonte de dados para a federação, exportar os dados que estarão disponíveis para a comunidade e realizar a equivalência de objetos. Quando uma fonte de dados é adicionada à federação a ferramenta realiza um teste para validar a conexão informando a disponibilidade da mesma. O fato de a fonte ter sido adicionada com sucesso não significa que ela está com os dados disponíveis para a federação.

Para disponibilizar os dados é necessário realizar a exportação dos esquemas locais, através da Tabela de Objetos Locais (TOL) e do Esquema de Exportação (EE). O usuário consulta os esquemas locais da fonte de dados e então seleciona aqueles que deseja exportar.

No momento que existirem pelo menos duas fontes de dados participando da federação e que possuam objetos presentes no esquema de exportação é possível realizar o mapeamento de equivalências entre as fontes. A ferramenta analisa os dados exportados, presentes no esquema de exportação, e propõe as equivalências através das seguintes regras: a) dois objetos com o mesmo nome da TOL em fontes diferentes são equivalentes; b) nomes de atributos iguais no EE e ambos com o mesmo *unique* são equivalentes ou nome de atributo de uma fonte igual ao nome de objeto de outra fonte são equivalentes ou nome de objeto de uma fonte seja igual ao nome de atributo de outra fonte são equivalentes; e c) papéis – O objeto não está no TEO e o papel está na TEO ou o objeto está na TEO e o papel não está na TEO ou o papel está na TOL e o papel não está na TEO e o objeto não está na TEO são equivalentes. Na figura 4 são apresentados exemplos de equivalências identificadas automaticamente pela ferramenta e que permitem ao usuário a confirmação da respectiva equivalência entre os objetos.

4. Conclusões

Este trabalho apresenta uma ferramenta web, aberta, para mapeamento de equivalências semânticas entre fontes de dados heterogêneas. A ferramenta, através de regras, identifica as possíveis equivalências entre objetos de fontes diferentes e apresenta ao usuário para validação.

Além disso, a ferramenta desenvolvida permite a vinculação dinâmica de qualquer fonte de dados que possua acesso através de *drivers JDBC*, para participação na federação, possibilitando uma interoperabilidade maior entre fontes de dados heterogêneas.

Como estudo de caso a ferramenta está sendo utilizada na Universidade Federal de Santa Maria (UFSM), como integradora de sistemas de diversos departamentos, envolvendo diferentes fabricantes de bancos de dados, como *IBM/DB2*, *Oracle* e *Firebird*. A ferramenta está disponível através do site <http://portal.ufsm.br/coridora>.

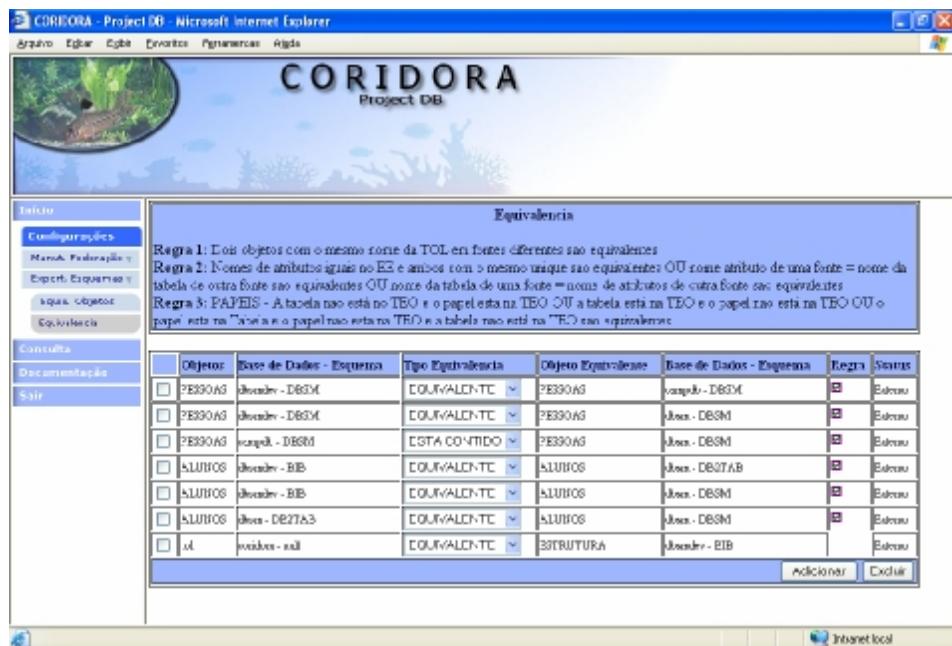


Figura 4. Identificação das Equivalências.

Os resultados alcançados com este artigo abrem a perspectiva de implementações de futuros trabalhos. Uma proposta é a extensão do modelo de mapeamento de esquemas para suportar XML *schemas*. Além da realização do mapeamento entre as fontes torna-se necessária a possibilidade de consulta de forma uniforme e complementar das informações exportadas. O desenvolvimento de ferramentas de consultas pode ser incluído neste contexto. Além das ferramentas de consulta, a implementação de algoritmos de similaridade também podem ser incluídos como trabalhos futuros, integrando e identificando divergências entre as diversas fontes participantes.

Referências

- DEITEL, Harver; DEITEL, Paul. **Java: Como programar.** 6^a ed. : Pearson, 2005.
- KANTORSKI, Gustavo Zanini. **Interoperabilidade de Bancos de Dados Heterogêneos Através da WWW.** Porto Alegre: PPGC da UFRGS, 1999. Dissertação de Mestrado.
- KANTORSKI, G. Z. ; RIBEIRO, C. H. F. . Heterogeneous Database Interoperability using the WWW. In: Simpósio Brasileiro de Banco de Dados, 2000, João Pessoa. XV Simpósio Brasileiro de Banco de Dados, 2000. p. 79-88.
- OZSU, M.; VALDURIEZ, P. Princípios de Sistemas de Bancos de Dados Distribuídos. Ney Jersey, 2001.
- RIBEIRO, Cora Helena Francisconi Pinto. **Banco de Dados Heterogêneos: Mapeamento dos Esquemas Conceituais em um Modelo Orientado a Objetos.** Porto Alegre: CPGCC da UFRGS, 1995. 165p. Tese de Doutorado.