Integração de Dados Heterogêneos em Ambiente Web

Álvaro C. P. Barbosa , Cristiano Biancardi , Leonardo Jose Silvestre

[1]Programa de Pós-Graduação em Informática - PPGI
Departamento de Informática
Universidade Federal do Espírito Santo
Campus de Goiabeiras
Av. Fernando Ferrari, S/N 29060-970 Vitória, ES

{alvaro,cbiancardi,lsilvestre}@inf.ufes.br

Resumo. *Vivemos em uma época na qual a chamada Sociedade da Informação necessita cada vez mais de acesso à informação, de forma confiável e rápida. O advento da Internet, aliado à globalização e à competitividade, fez crescer dramaticamente tanto a oferta de novas informações quanto a demanda por informação já disponível. Para tanto, sistemas de integração de dados vêm sendo pesquisados e desenvolvidos, na tentativa de prover aos usuários uma visão única, uniforme e homogênea, a partir de diversas fontes de dados heterogêneas, distribuídas e desenvolvidas independentemente. O objetivo principal desses sistemas é prover transparência de localização, distribuição e heterogeneidade dos dados.*

Abstract. *should not have more than 10 lines and must be in the first page of the paper. Vivemos em uma época na qual a chamada Sociedade da Informação necessita cada vez mais de acesso à informação, de forma confiável e rápida. O advento da Internet, aliado à globalização e à competitividade, fez crescer dramaticamente tanto a oferta de novas informações quanto a demanda por informação já disponível. Para tanto, sistemas de integração de dados vêm sendo pesquisados e desenvolvidos, na tentativa de prover aos usuários uma visão única, uniforme e homogênea, a partir de diversas fontes de dados heterogêneas, distribuídas e desenvolvidas independentemente. O objetivo principal desses sistemas é prover transparência de localização, distribuição e heterogeneidade dos dados.*

1. Introdução

Historicamente, as organizações desenvolveram múltiplas ilhas de sistemas de computadores e de bancos de dados, com o propósito de atender seus requisitos específicos. O aumento e a globalização dos negócios, a fusão e parcerias entre organizações, a existência de uma competitividade agressiva e o advento da Internet, fizeram crescer dramaticamente tanto a oferta de novas informações quanto a demanda por informação já disponível. Neste contexto, o número de aplicações que requerem acesso integrado a múltiplas fontes de dados, heterogêneas e distribuídas, cresceu muito e surgiu a necessidade da pesquisa e o desenvolvimento de soluções que atuassem como pontes para integração destas ilhas de informações. O acesso a tais informações deve ser realizado através das redes de computadores organizacionais, das intranets corporativas e da Internet [Park, 2001].

Os sistemas de integração de dados têm como objetivo fornecer aos usuários uma interface uniforme para as diversas fontes de dados, possivelmente autônomas, heterogêneas e distribuídas [Özsu and Valduriez, 2001]. Ou seja, o objetivo é liberar o

usuário de ter que localizar as fontes de dados, interagir com cada uma isoladamente e combinar manualmente dados vindos dessas múltiplas fontes [Halevy, 2003].

Tradicionalmente, as fontes de dados são bancos de dados ou arquivos de dados de sistemas legados. Recentemente, encontramos novas fontes de dados, dentre as quais figuram com destaque as páginas *Web* que, em geral, são semi-estruturadas. Através da *Web*, qualquer indivíduo ou organização pode se tornar um fornecedor de informação sem requerer autorização, ou seja, a quantidade de informação disponível aumentou (e continua aumentando) consideravelmente.

Uma grande quantidade de informações e serviços, heterogêneos e distribuídos (por exemplo, home pages, bibliotecas digitais *online*, catálogos de produtos etc.), está prontamente disponível [Bouguettaya et al., 2000], e os usuários, cada vez mais, necessitam de uma visão integrada dos dados disponíveis a partir dessas fontes de dados [Hakimpour and Geppert, 2001].

Desde o surgimento dos sistemas de gerenciamento de banco de dados, o pro-blema de integração de dados tem sido reconhecido como ubíquo e criticamente importante [Miller et al., 2001], e vem demandando pesquisa na área de Banco de Dados por mais de uma década [Sheth and Larson, 1990, Litwin et al., 1990, Silberschatz and Zdonik, 1997, Abiteboul et al., 2003, Halevy, 2003]. Inicialmente, a preocupação era integrar dados de diferentes sistemas em uma mesma organização (por exemplo, integrar dados dos sistemas contábil e administrativo). Com a globalização, e as conseqüentes aquisições de empresas por outras, fusões e acordos comerciais entre diferentes organizações, o problema passou a ser integrar dados de organizações distintas. Posteriormente, com o advento da *Web*, surgiu a necessidade de integrar dados dos mais diferentes tipos, seja por organizações, seja por indivíduos. Através da Internet, organizações podem compartilhar informações e fazer negócios (por exemplo, sistemas B2B e B2C - *Business to Business* e *Business to Consumer*, respectivamente); instituições governamentais e de pesquisa podem compartilhar dados entre si; e mesmo usuários individuais precisam de algum tipo de acesso uniforme à grande quantidade de informação disponível na *Web*.

O restante do texto é organizado como segue. A seção 2 apresenta o cenário atual da integração de dados. A seção 3 apresenta conceitos básicos para melhor entendimento da integração de dados, e algumas tecnologias relacionadas com a área. Na seção 4 são apresentadas algumas das soluções para integração de dados existentes, divididas em sistemas acadêmicos e sistemas comerciais. A seção 5 apresenta problemas existentes e desafios a serem superados. Finalmente, a seção 6 conclui, fazendo algumas considerações.

2. Cenário Atual

A disseminação da Internet nos últimos anos fez crescer enormemente tanto a disponibilidade quanto a demanda por informação. Tanta informação disponível trouxe à tona uma necessidade: como descobrir, acessar e obter uma visão integrada dessas informações, que são heterogêneas, por terem sido disponibilizadas independentemente por indivíduos e organizações diferentes; e distribuídas, pelo simples fato de estarem na Internet?

As máquinas de busca são uma tentativa de se prover uma visão integrada inicial, por assunto, dentre a grande quantidade de informação disseminada na *Web*. Através

de um *site Web* (por exemplo, *Google*[1], *Altavista*[2], *Cadê*[3]), uma pessoa pode pesquisar por informações, obtendo como resultado um conjunto de páginas que contém e/ou estão relacionadas com o assunto procurado. A qualidade das informações retornadas é muitas vezes questionável, até por que as elas nem sempre estão atualizadas, mas é fato que as máquinas de busca são ferramentas extremamente úteis quando se navega pela Internet.

Em se tratando de organizações, uma técnica muito usada para integração de dados é a de *Data Warehousing*. Um *Data Warehouse* é um conjunto de dados baseado em assuntos, integrado, não volátil, e variável em relação ao tempo, de apoio às decisões gerenciais [Inmon, 1997]. Isto é, um banco de dados de suporte a decisão que pode ser extraído e consolidado a partir de um conjunto de fontes de dados, ou seja, é um tipo de integração de diversas fontes em uma só, de forma que se possa extrair informações (por exemplo, relatórios gerenciais) que envolvem dados de diversos setores de uma organização.

Integração de Aplicações Empresariais (*Enterprise Application Integration* (EAI), tal como a *Data Warehousing*, envolve a captura e a transformação de dados, mas o faz de uma forma diferente. Durante o desenvolvimento da organização, diversos sistemas foram criados para atender a necessidades específicas (por exemplo, sistema de folha de pagamentos, sistema contábil, sistema financeiro etc.). O objetivo da EAI é oferecer uma forma integrada de se lidar com todos esses sistemas. A tecnologia EAI possibilita um intercâmbio de informações à medida que ocorrem eventos, e não a transferência periódica de arquivos *batch*, como ocorre no *Data Warehousing* [Cummins, 2002].

A Internet possibilitou o estabelecimento de relações comprador/vendedor antes inimagináveis. As distâncias foram incrivelmente encurtadas, possibilitando o surgimento do *e-commerce*, ou comércio eletrônico, que é ”basicamente a integração corporativa que se prolonga para clientes e parceiros de negócio”[Cummins, 2002]. Ele aparece em duas formas principais: B2C (comércio eletrônico empresa/cliente) e B2B (comércio eletrônico empresa/empresa), e envolve diretamente a integração de dados. Além do *e-commerce*, ganharam força o *e-learning*, ou aprendizado eletrônico, que envolve educação à distância, treinamento baseado na *Web* entre outros; e a *e-science*, que pode ser definida como ”a ciência feita através de colaborações globais, habilitadas pela Internet, usando gigantescas coleções de dados, recursos de computação em alta escala e visualização de alta performance”[Rickett, 2001].

Recentemente, especialmente no meio acadêmico, muito tem se pesquisado sobre a Web Semântica (*Semantic Web*). Segundo Tim Berners-Lee, criador da *World Wide Web* [W3C, 2004b], ela é uma extensão da *Web* tradicional, na qual, a partir do uso intensivo de metadados, espera-se obter o acesso automatizado às informações, com base no processamento semântico de dados e heurísticas feitos por máquinas. A Web Semântica se relaciona fortemente com a Integração de Dados, tendo em vista que existe a necessidade de se integrar os serviços e dados relacionados à ela. Além disso, a Web Semântica pode auxiliar as máquinas de busca, visto que possibilita um melhor entendimento das informações que estão disponíveis, facilitando o refinamento das consultas.

Além disso, diversas aplicações ganharam destaque ultimamente no que se refere à integração de dados. Dentre elas, podemos destacar as ambientais, as médicas e as industriais, especialmente aquelas que envolvem dados de sensores, os quais produzem uma grande quantidade de dados que são, em geral, heterogêneos.

[1] http://www.google.com.br
[2] http://www.altavista.com
[3] http://www.cade.com.br

3. Conceitos Básicos e Tecnologias Relacionadas

Para melhor entendimento sobre o assunto, são apresentados a seguir alguns conceitos básicos e tecnologias relacionadas com integração de dados.

Bancos de Dados Federados [Sheth and Larson, 1990]: Coleção de sistemas de bancos de dados (SBDs) componentes que cooperam entre si, mas são autônomos. Os bancos de dados componentes são integrados em vários graus. O software que provê a manipulação controlada e coordenada dos bancos de dados componentes é chamado *Federated Database Management System - FDBMS* (Sistema de Gerenciamento de Bancos de Dados Federados - SGBDF). A figura 1 apresenta um SBDF (Sistema de Bancos de Dados Federados) e seus componentes.

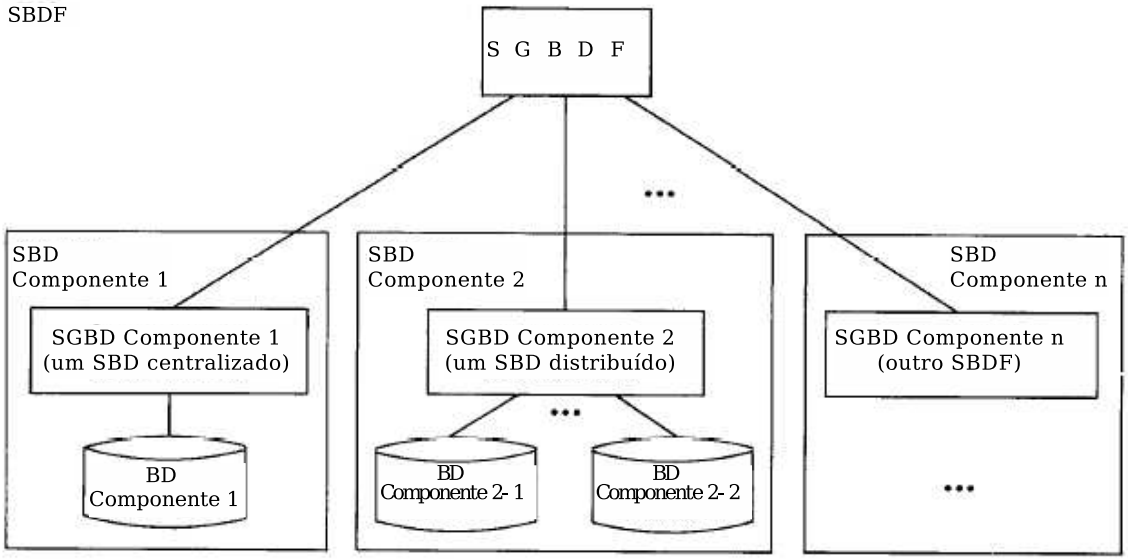

Figura 1: Um SBDF e seus componentes [Sheth and Larson, 1990].

Banco de Dados x Fonte de Dados: Inicialmente, o objetivo maior dos sistemas integradores era integrar bancos de dados. Mais recentemente, tais sistemas têm como objetivo integrar os mais diversos tipos de fontes de dados, as quais podem ser arquivos-texto, páginas HTML, dentre outros, inclusive os próprios bancos de dados.

Esquemas Local, de Exportação e Global: Os esquemas locais são os esquemas de dados que cada fonte possui. Os de exportação são aqueles que a fonte deseja disponibilizar para o sistema integrador. Já o global é aquele que representa o esquema integrado de todas as fontes, possuindo mapeamentos que indicam as correspondências entre os atributos globais e os atributos de exportação. Os usuários do sistema integrado fazem consultas levando em consideração o esquema global. A figura 2 apresenta os tipos de esquemas e suas relações.

Modelo de Dados Comum: Modelo de dados no qual os dados do sistema integrador serão representados, ou seja, o modelo para o qual devem ser convertidos os modelos de dados heterogêneos que se deseja integrar. Por exemplo, podemos ter modelos de dados comuns XML, Relacional, Orientado a Objetos, baseado em Ontologias etc. Também é conhecido como Modelo de Dados Canônico. O esquema global (figura 2) é representado através do modelo de dados comum.

Abordagens tradicionais para projeto de um sistema de integração de dados:

- ***Global-As-View* - GAV**: Os esquemas globais são definidos como visões sobre as fontes de dados, ou seja, constrói-se o esquema global a partir de visões sobre os esquemas de exportação das fontes.

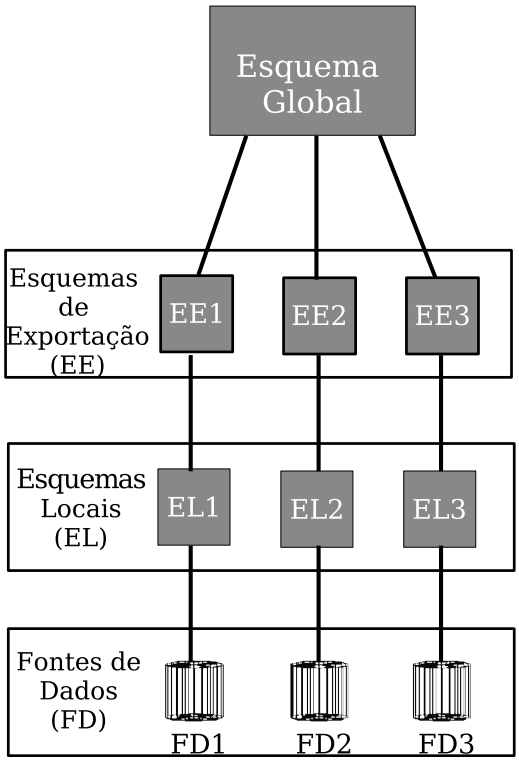

Figura 2: Tipos de Esquemas e suas relações.

- ***Local-As-View* - LAV**: Os esquemas das fontes são definidos como visões sobre o esquema global, ou seja, constrói-se primeiro o esquema global e, a partir dele, visões que representam os esquemas das fontes.

Interoperabilidade: Capacidade de compartilhar e trocar informações e processos entre ambientes computacionais heterogêneos, autônomos e distribuídos [Yuan, 1998].

Wrappers: Programas usados para fazer a tradução entre o modelo de dados comum, utilizado no sistema integrador, e o modelo de dados de cada fonte. Por exemplo, se o modelo de dados comum é relacional e queremos integrar fontes de dados XML ou Orientadas a Objetos - OO, temos que construir *wrappers* que traduzam de XML ou OO para Relacional. Além disso, os *wrappers* são utilizados para prover a comunicação com as fontes de dados. A figura 3 mostra *wrappers* em um sistema integrador.

Dados Estruturados Apresentam estrutura de representação ou esquema previamente definido e bem comportado. Exemplo: bancos de dados relacionais.

Dados Semi-Estruturados: Apresentam uma representação estrutural heterogênea. Exemplo: páginas HTML, arquivos XML.

Dados Não-Estruturados: Não apresentam estrutura definida. Exemplos: texto livre, vídeos, figuras.

Integração de Informações na Web x Integração de Informações em Bancos de Dados [Hsu et al., 2003]: Integração de informações na *Web* é diferente de integração de informações em bancos de dados devido à natureza da *Web*, na qual os dados estão contidos em páginas interligadas no lugar de tabelas ou objetos com um esquema claramente definido como em sistemas de bancos de dados. Um dos grandes problemas da integração de dados na *Web* é a natureza ”sem estado”(*stateless*) das transações *Web*, em decorrência do protocolo HTTP usado na comunicação cliente *Web* e servidor *Web* [de Lima, 1997], diferente dos SGBDs, nos quais o estado das informações sobre as ações do usuário é mantido, enquanto durar sua interação com o SGBD (*statefull*).

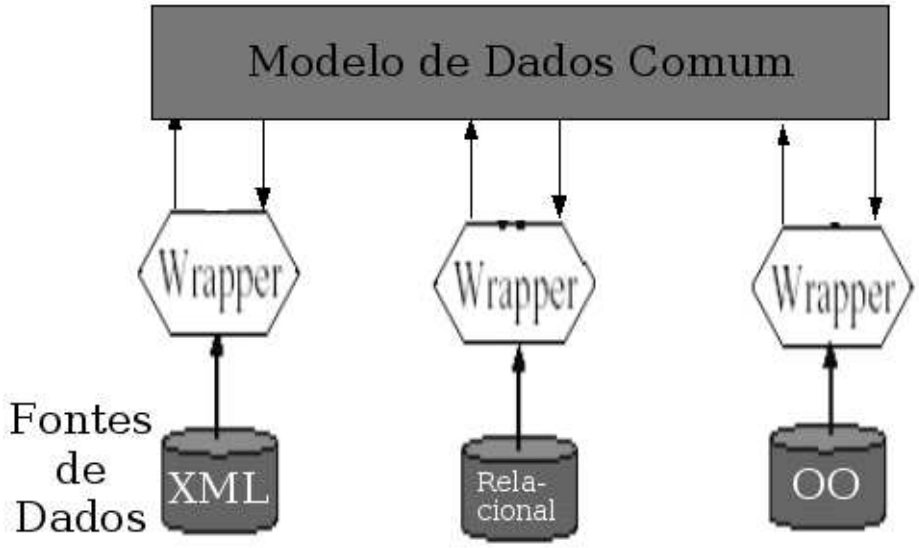

Figura 3: Wrappers em um sistema integrador.

XML: XML (*eXtensible Markup Language*) [W3C, 2004a] é uma linguagem desenvolvida para a descrição de dados (conteúdo) que permite a criação de formatos únicos para descrever dados de aplicações específicas. XML vem se tornando rapidamente um padrão para representação e troca de dados. Assim, ela pode ajudar na integração de dados estruturados, semi-estruturados e não-estruturados na *Web*, sendo bastante adequado para ser usada como um modelo de dados comum na *Web* [Bertino and Ferrari, 2001].

Ontologias: Recentemente, têm sido dada uma maior ênfase ao uso de ontologias na integração de dados [Wache et al., 2001], já que elas possibilitam um maior entendimento do domínio, por permitirem a definição de um vocabulário padrão, orientado a domínio. Na Ciência da Computação, ontologias geralmente são definidas como uma especificação formal de uma conceitualização que é senso comum para um grupo de pessoas [Gruber, 1993]. Elas podem, então, ser utilizadas como modelos conceituais que provêm suporte semântico para incrementar a manipulação e o mapeamento de dados [Mena et al., 1996, Embley et al., 1998, Bergamaschi et al., 1999, Martin and Eklund, 1999, Erdmann and Studer, 2001].

Metadados: Metadados são ”dados sobre dados”, ou seja, aqueles dados que descrevem os dados armazenados. Dentre as formas de representação de metadados (dados sobre os dados) que têm sido propostas, podemos destacar aquelas que se relacionam com ontologias, em especial RDF (*Resource Description Framework*) [Lassila and Swick, 1998], que é uma recomendação da W3C (*World Wide Web Consortium*), e pode ser usado como um *framework* geral para troca de dados na *Web* [Decker et al., 2000].

Web Services: Segundo David Smith, do *Gartner Group*, ”*Web Services* são conteúdo e processos de software distribuídos pela Internet, usando mensagens fracamente acopladas (e interfaces XML) que atendem a um conjunto particular de necessidades de usuários”. A tecnologia de *Web Services* surgiu da necessidade de comunicação e cooperação entre aplicações, especialmente na internet. Em organizações com aplicações heterogêneas e arquiteturas distribuídas, a introdução de Web Services padroniza a comunicação e propicia a interoperabilidade de aplicativos escritos em diferentes linguagens de programação residindo em diferentes plataformas [Kreger, 2001].

4. Soluções Existentes

Esta seção apresenta alguns dos principais sistemas existentes para integração de dados, divididos em duas categorias: Sistemas Acadêmicos e Sistemas Comerciais[4].

Diversos tipos de soluções para o problema de integração de dados têm sido propostos. Historicamente, podemos destacar os Sistemas de Gerenciamento de Bancos de Dados Heterogêneos (SGBDH) [Litwin et al., 1990, Sheth and Larson, 1990, Thomas et al., 1990, Pitoura et al., 1995, Silberschatz and Zdonik, 1997, Conrad et al., 1997] e os Sistemas Mediadores [Wiederhold and Genesereth, 1997, Wiederhold, 1998]. Mais recentemente, ambos têm sido denominados *middleware* para integração de dados [Barbosa, 2001, Haas et al., 99, Rodriguez-Martinez and Roussopoulos, 2000]. *Middleware* é um componente de software que tem como finalidade interligar processos clientes a processos servidores, disponibilizando um conjunto de serviços e visando a reduzir a complexidade do processo de desenvolvimento de uma aplicação. *Middleware* pode ser visto como uma ferramenta que "salva os desenvolvedores de se envolver em detalhes de cada fonte de dados"[Linthicum, 1997].

4.1. Sistemas Acadêmicos

- **Le Select**
  Le Select [Xhumari et al., 2000] é um middleware desenvolvido no INRIA (França) através da abordagem de mediadores, cujo principal objetivo é disponibilizar dados e programas, além de permitir a integração de fontes de dados heterogêneas. O Le Select possui uma arquitetura distribuída onde os servidores se comunicam segundo o modelo *peer-to-peer* (figura 4) e a integração de dados é realizada através do modelo relacional. Um site publicador disponibiliza os seus dados e/ou programas para qualquer cliente Le Select via *Web*.

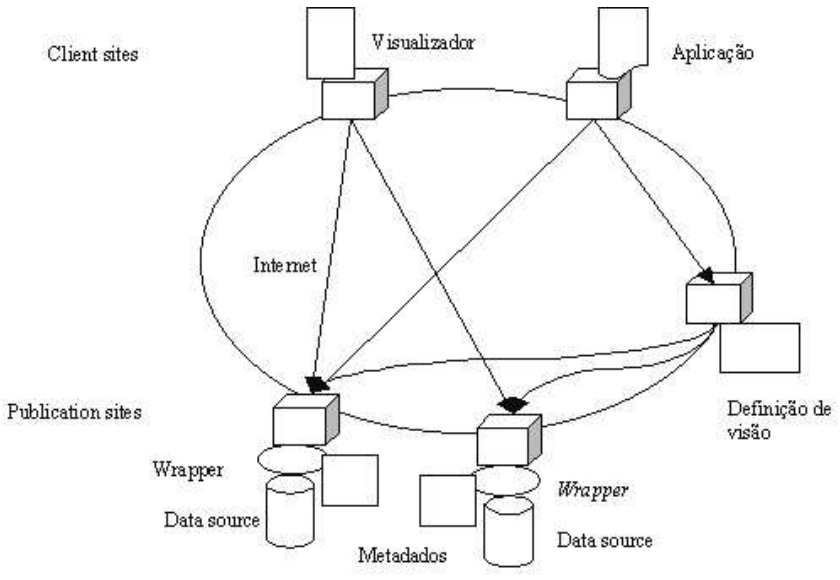

Figura 4: Arquitetura do Le Select [Xhumari et al., 2000].

No Le Select, os clientes podem acessar os dados através de consultas SQL e podem disparar a execução de programas, de forma assíncrona, nos sites publicadores de programas que geram sessões persistentes de execução.

[4]É importante ressaltar que os sistemas comerciais foram influenciados diretamente por resultados de pesquisas. A vasta maioria dos produtos de integração de dados foram desenvolvidos por membros da comunidade de pesquisa, seja na forma de criação de companhias, seja por comercialização de tecnologias a partir de laboratórios de pesquisa industrial [Halevy, 2003]

- **Agora**
  O Agora é desenvolvido sobre o Le Select, tendo como principal adicional o fato de empregar XML como formato de interface do usuário, enquanto todos os fluxos de dados no processador de consultas consistem de tuplas relacionais, ou seja, o modelo interno é relacional e a máquina relacional é transparente para o usuário, que enxerga tudo como se fosse XML [Manolescu et al., 2000]. Dessa forma, as consultas são feitas utilizando-se XML (Xquery), e o resultado é retornado para o usuário em forma de documentos XML. A figura 5 apresenta a arquitetura do Agora e sua relação com o Le Select.

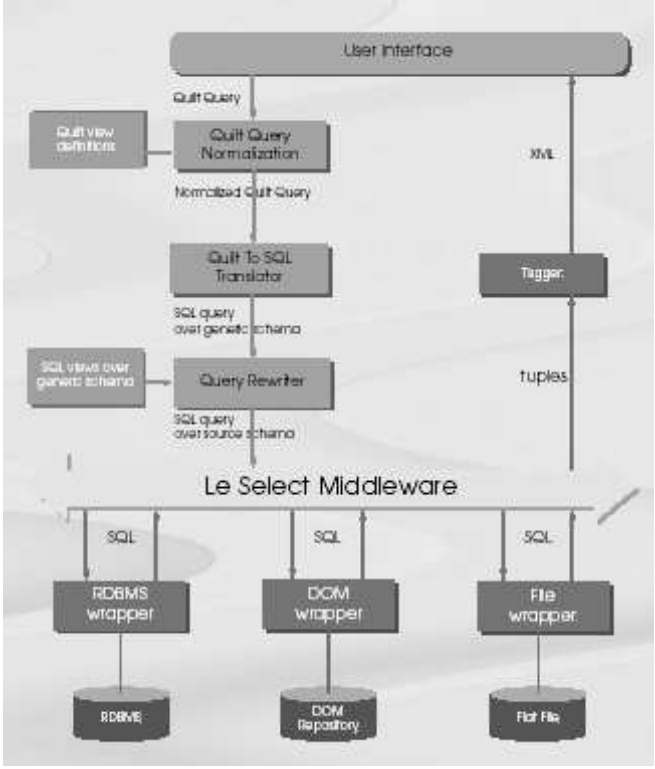

Figura 5: Arquitetura do Agora sobre o Le Select [Florescu et al., 2000].

- **MOCHA**
  MOCHA [Rodriguez-Martinez and Roussopoulos, 2000] (Middleware Based on a Code Shipping Architecture) é um middleware proposto pela Universidade de *Maryland*, baseado em serviços de banco de dados. Foi projetado para integrar centenas de fontes de dados distribuídas sobre a *Web*, tendo sido construído com a idéia de que um middleware para um ambiente distribuído de larga escala deve ser auto-extensível. Esta extensibilidade permite o envio de classes Java para os sites remotos de um modo automático, de acordo com as características das fontes de dados. As classes Java são necessárias para executar uma sub-consulta no próprio site, diminuindo o tráfego de dados entre o site e o sistema. A figura 6 apresenta a arquitetura do MOCHA.
- **MOMIS/MIKS**
  MOMIS (*Mediator envirOnment for Multiple Information Sources*) [Beneventano et al., 2001a, Bergamaschi et al., 1999] é um *framework* para executar integração e extração de informação de fontes de dados estruturadas e semi-estruturadas, fornecendo acesso integrado para informações heterogêneas armazenadas em bases de dados tradicionais (por exemplo, relacional, orientado a objeto) ou sistemas de arquivos, como também a fontes semi-estruturadas. É desenvolvido pelas universidades de Modena, Reggio Emilia, Milano e Brescia.

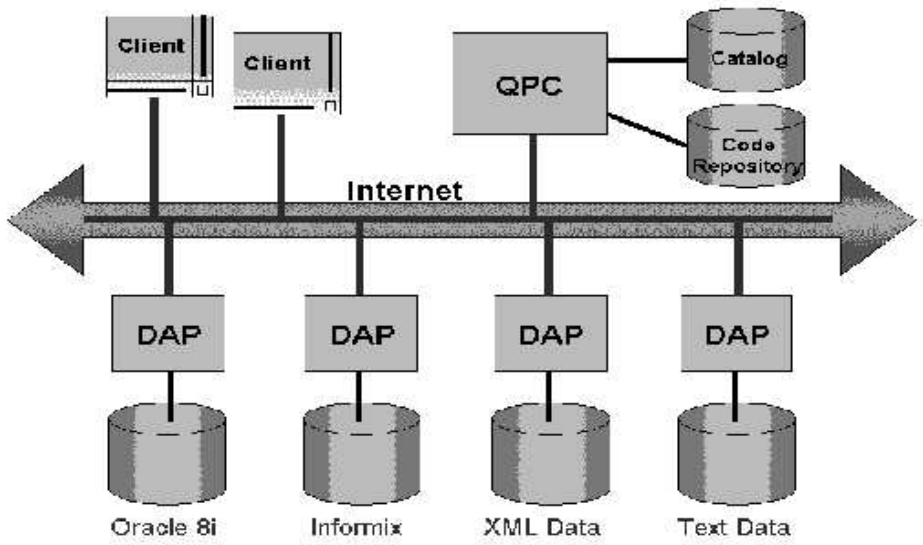

Figura 6: Arquitetura do MOCHA [Rodriguez-Martinez and Roussopoulos, 2000].

Para que se possa fazer a integração de informações, MOMIS, tal como outros projetos de integração [Arens et al., 1993, Roth and Schwarz, 1997], permite uma ”abordagem semântica”baseada no esquema conceitual, ou metadado, das informações das fontes. A figura 7 apresenta a arquitetura do MOMIS.

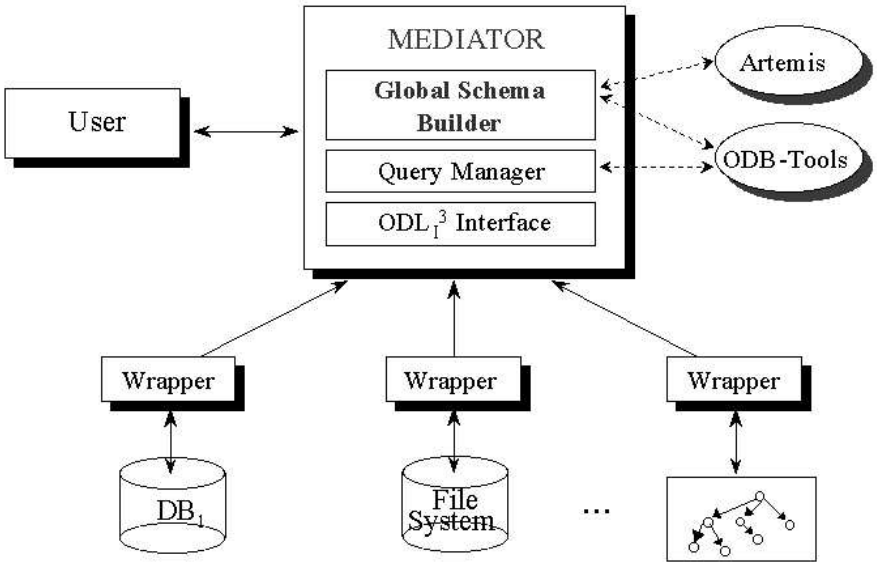

Figura 7: Arquitetura do MOMIS [Beneventano et al., 2001b].

O MIKS [Bergamaschi et al., 2001b] (*Mediator agent for Integration of Knowledge Sources*) enriquece a arquitetura do MOMIS explorando as características de agentes móveis e inteligentes. Ele executa extração, manipulação e consulta de informação através de agentes de software integrados ao processo de integração de informação do MOMIS. A figura 8 apresenta a arquitetura do MIKS.

- **IBIS**
  IBIS [Calì et al., 2003, Calì et al., 2002] (Internet-Based Information System) é um sistema para integração semântica de fontes de dados heterogêneas, o qual adota soluções inovadas para lidar com todos os apectos de um complexo ambiente de integração de dados incluindo o uso de *wrappers* para as fontes, limitações no acesso às mesmas, reposta a consultas sob restrições de integridade. Este sistema

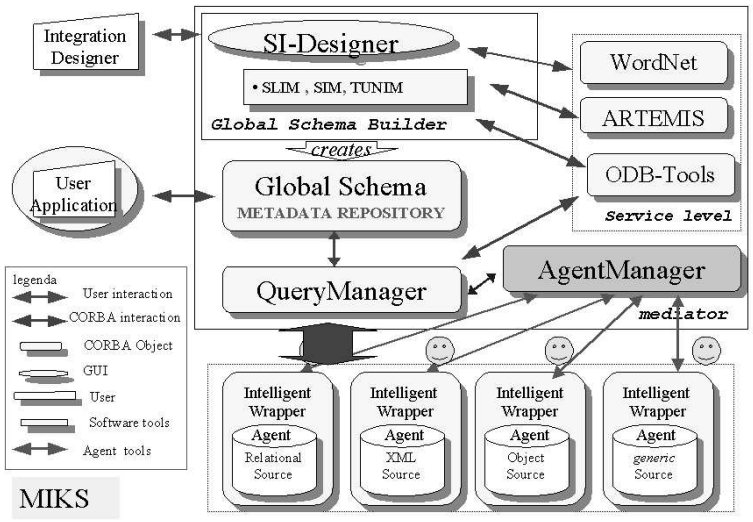

Figura 8: Arquitetura do MIKS [Bergamaschi et al., 2001a].

foi estudado e desenvolvido em colaboração entre a Universidade de Roma ”La Sapienza”e CM Sistemi.

O IBIS é baseado na abordagem de *global-as-view* (GAV) usando um esquema mediado relacional para consultar os dados nas fontes. O sistema é capaz de lidar com uma variedade de fontes de dados, incluindo fontes de dados na Web, banco de dados relacionais e fontes legadas. Cada fonte não relacional é traduzida para fornecer uma visão relacional da mesma. Cada fonte é considerada incompleta, no sentido que ela contribui com dados para o sistema de integração.

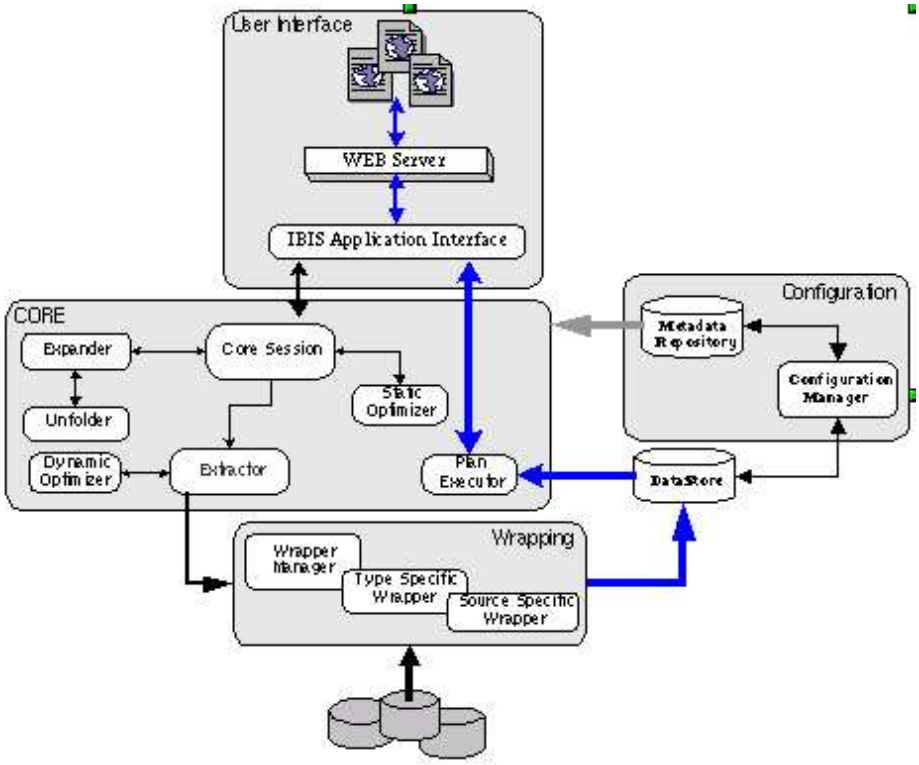

Figura 9: Arquitetura do MIKS [Calì et al., 2003].

- **CoDIMS**
  O CoDIMS [Barbosa, 2001, Barbosa et al., 2002, Trevisol, 2004, Fontes et al., 2004] é um ambiente *middleware* para a geração de sistemas *middleware* flexíveis e configuráveis para integração de dados, que vem sendo desenvolvido pelo Grupo de Pesquisa em Banco de Dados do Departamento de Informática da Universidade Federal do Espírito Santo - UFES e instituições parceiras. Por ser flexível, o CoDIMS tem como propósito permitir integrar diferentes fontes de dados; utilizar diferentes formas de comunicação; utilizar diferentes modelos de dados como modelo integrador; e utilizar diferentes linguagens de consulta, de acordo com os requisitos da aplicação. Também de acordo com as necessidades da aplicação, pelo fato de o CoDIMS ser configurável, pode-se escolher componentes necessários e adequados a serem incluídos em um sistema configurado específico. Por exemplo, uma aplicação que necessite somente de fazer consultas, mas não atualizações, poderia não utilizar componentes como Gerência de Transação e Controle de Concorrência. A figura 10 mostra o CoDIMS e seus componentes.

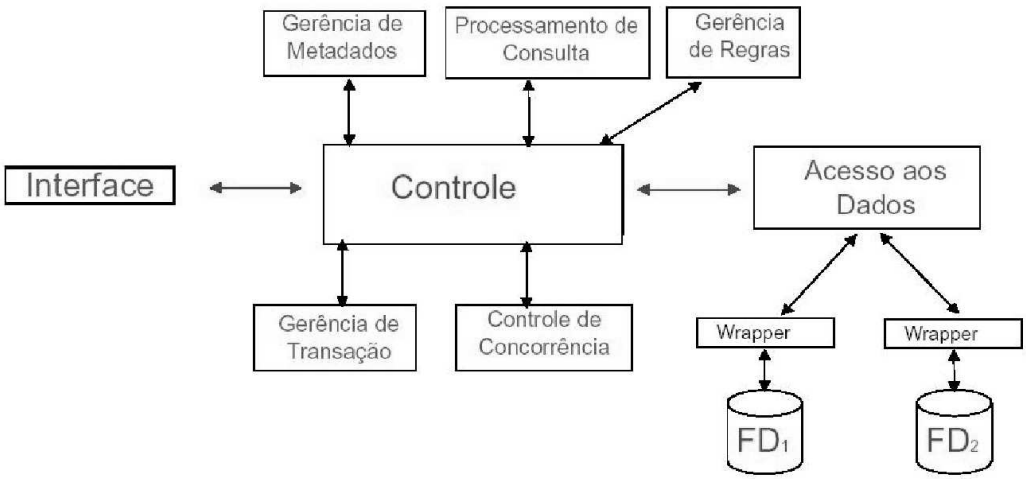

Figura 10: O CoDIMS e seus componentes.

4.2. Sistemas Comerciais

- **DB2ii**
  O DB2ii (DB2 *Information Integrator*), desenvolvido pela IBM®, é um produto para integração de informações de fontes de dados heterogêneas. O DB2ii possui capacidade de federar, pesquisar, fazer cache, transformar e replicar dados [IBM, 2003]. Ele integra dados de fontes variadas, tais como bancos de dados relacionais, documentos XML, planilhas do Microsoft® Excel, entre outras. Além disso, o DB2ii oferece capacidade de replicação de dados, através de um servidor de replicação.
  No DB2, um sistema federado consiste de uma instância do DB2 que opera como servidor federado, um banco de dados que atua como o banco de dados federado, uma ou mais fontes de dados, e clientes (usuários e aplicações) que acessam o banco e as fontes de dados [IBM, 2002].
  As consultas são feitas através de SQL, e os resultados podem ser retornados como conjuntos de respostas SQL ou documentos XML. Além das consultas so-

bre a federação, é possível, através de um modo especial chamado *pass-through* [IBM, 2002], submeter consultas SQL diretamente às fontes de dados.
No seu catálogo global, o DB2ii armazena informações sobre as fontes de dados, tais como aquelas que o servidor federado usa para conectar com essas fontes e os mapeamentos entre as autorizações do usuário federado para as autorizações do usuário da fonte de dados [IBM, 2002]. A figura 11 mostra a arquitetura do DB2ii, com os componentes de um sistema federado do mesmo e algumas das fontes de dados suportadas.

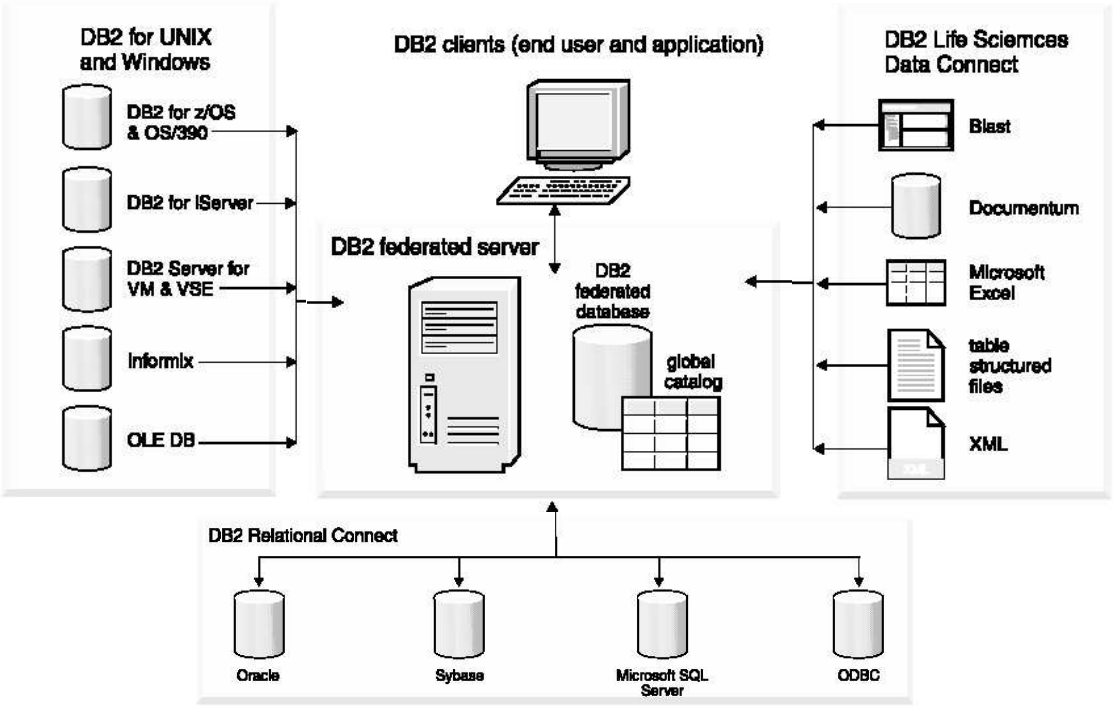

Figura 11: Arquitetura do DB2ii [IBM, 2003].

- **Denodo** [Pan et al., 2002]
  Sistema desenvolvido pela *Denodo Corporation*, que segue a arquitetura de mediador, cujos componentes e seus relacionamentos são apresentados na figura 12. A camada física compreende *wrappers* para diferentes fontes de dados: bancos de dados relacionais, sites *web*, *flat files*, planilhas eletrônicas etc. *Wrappers* são gerados semi-automaticamente por uma ferramenta. A camada lógica, chamada *Denodo Aggregation Engine* (Máquina de Agregação Denodo), é o núcleo da plataforma, e contém o Gerador de Planos de Consulta, o Otimizador e a Máquina de Execução. Uma ferramenta de administração é usada para gerenciar o dicionário de dados. Um módulo *cache* armazena as visões materializadas, evitando consultas nas fontes quando elas podem ser resolvidas utilizando a *cache*. A *Denodo Planning Tool* (Ferramenta de Planejamento Denodo) permite *data prefetches* periódicos. O *Denodo Security SDK* permite codificação de dados quando requerido.
- **Oracle**
  A *Oracle Corporation* oferece uma solução para integração de dados, com o banco de dados *Oracle9i*. Através da *Distributed SQL* (SQL distribuída) do *Oracle9i*, permite integração síncrona, consolidando as informações em tempo real, e escondendo a localização dos objetos da aplicação ou do usuário, de forma que os objetos nas fontes de dados remotas aparecem como locais. A integração assíncrona usa tecnologias de mensagem e replicação para mover dados de bancos de dados remotos para um banco de dados local, no qual as aplicações podem acessar

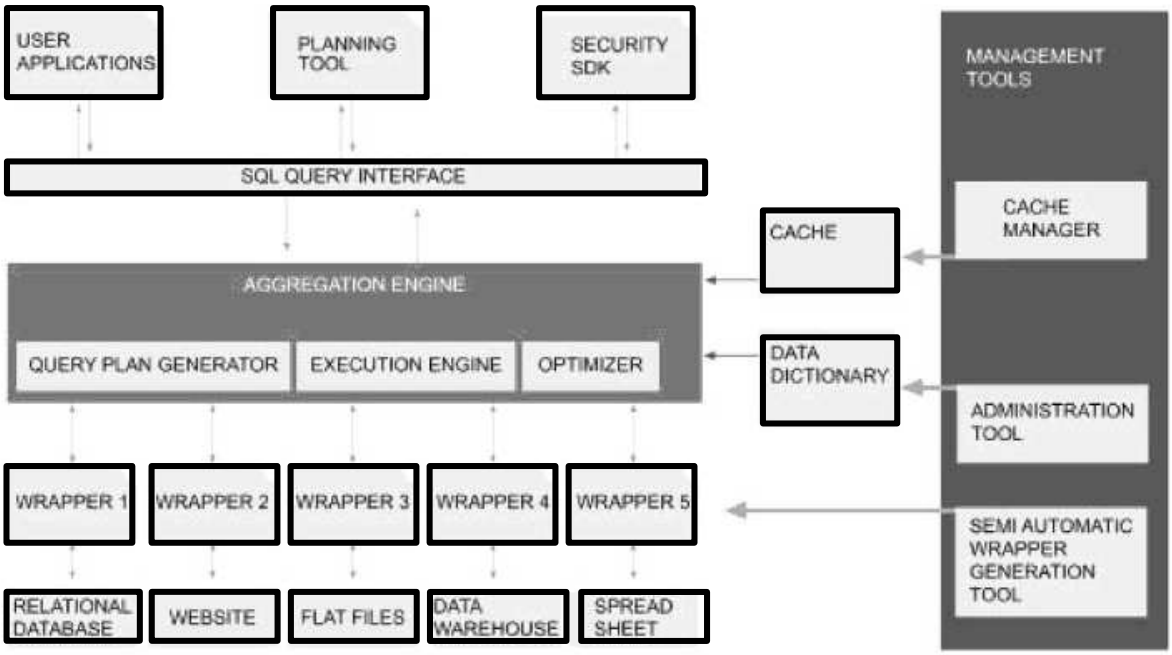

Figura 12: Arquitetura do Sistema Mediador Denodo [Pan et al., 2002].

diretamente os dados. Essa ferramenta, assim como o DB2ii, consegue integrar um conjunto pré-determinado de SGBDs [Oracle, 2003]. A figura 13 apresenta a plataforma de integração *Oracle9i*.

5. Problemas/Desafios

Integração de dados ainda é uma área de pesquisa bastante atual, na qual existem diversos problemas não resolvidos e desafios a serem enfrentados, dentre os quais podemos destacar:

Arquiteturas Flexíveis e Configuráveis: É necessário procurar abrir as arquiteturas de banco de dados onde novos serviços possam ser incorporados e permitir configurar a funcionalidade de maneira mais flexível, de acordo com as necessidades da aplicação [Silberschatz and Zdonik, 1997]. Assim, as arquiteturas flexíveis e configuráveis para integração de dados, tais como o CoDIMS [Barbosa, 2001, Barbosa et al., 2002, Trevisol, 2004], são de grande importância nessa área, tendo em vista que elas permitem gerar diferentes sistemas integradores, de acordo com as necessidades das mais diversas aplicações. É necessário, nesse contexto, consolidar implementações de tais arquiteturas, de forma que elas possam ser efetivamente utilizadas.

Representação, Mapeamento e Integração de Esquemas: A integração de esquemas, que consiste construir uma visão global de um conjunto de esquemas desenvolvidos independentemente [Sheth and Larson, 1990, Batini et al., 1986, Elmagarmid and Pu, 1990, Parent and Spaccapietra, 1998], é um problema que vem sendo pesquisado desde o início dos anos 80, sendo um dos principais motivadores dos trabalhos em *schema matching* [Rahm and Bernstein, 2001], e envolve diretamente a representação e o mapeamento dos esquemas das fontes, constituindo ainda hoje um desafio na área de integração de dados. O problema de integração de esquemas envolve principalmente combinar diferentes atributos, metadados e relacionamentos de diferentes fontes.

Conversão entre Modelos de Dados: O sistemas de integração existentes, em geral, possuem um modelo de dados único, ou seja, quando tal sistema utiliza o modelo rela-

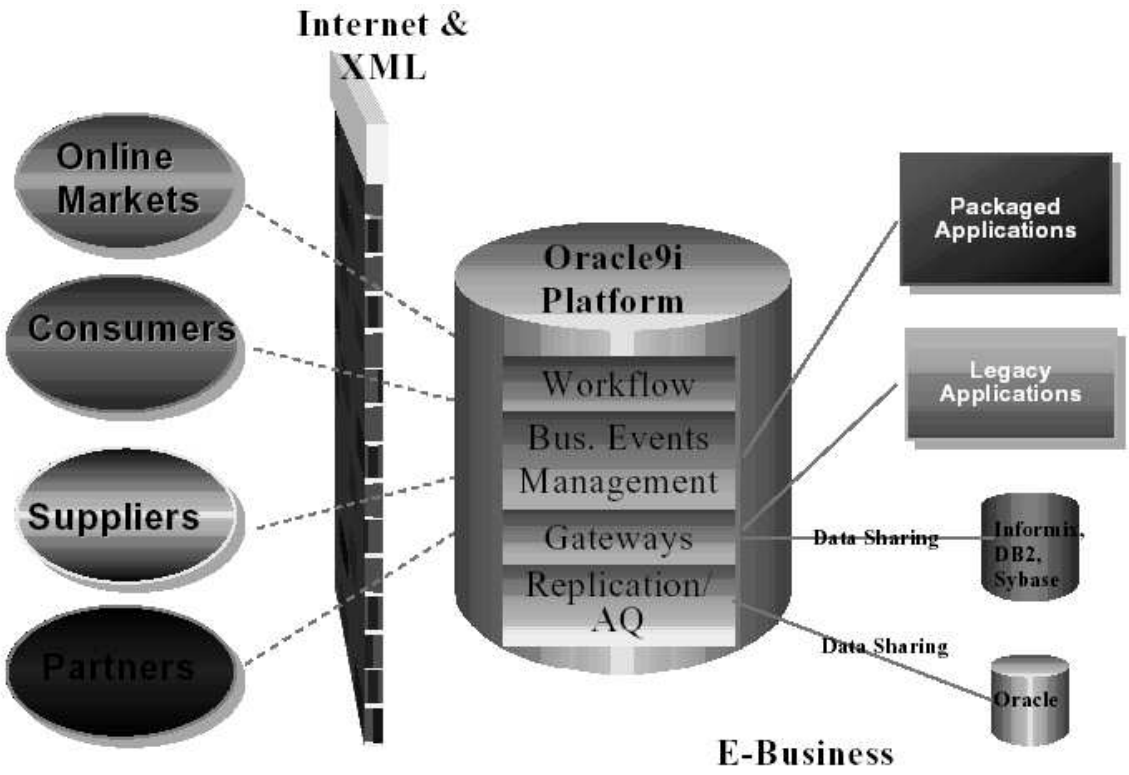

Figura 13: Plataforma de Integração *Oracle9i* [Oracle, 2000].

cional como modelo de dados, os usuários devem realizar consultas utilizando SQL. Um desafio importante é oferecer ao usuário possibilidades de utilizar diferentes modelos de dados, independentemente do modelo interno que o sistema utilize. É importante que sejam desenvolvidas técnicas para facilitar a criação de *wrappers*, os quais são os principais responsáveis pela conversão entre modelos.

Gerência de Inconsistência e Incerteza: quando se integra dados de múltiplas fontes de dados, os mesmos podem ser inconsistentes ou incertos. Dessa forma, é necessário desenvolver métodos para gerenciar tais dados, explicando as inconsistências e incertezas para o usuário, guiando-os através do processo de reconciliar diferentes conjuntos de dados [Halevy, 2003].

Resultados Imprecisos para Consultas: os sistemas de integração existentes enviam sub-consultas para cada fonte de dados que pode ter dados relevantes para responder às consultas, dando assim uma resposta completa para cada uma delas. Na *Web*, isso é inviável, já que, por exemplo, algumas fontes podem não responder, outras podem estar com dados desatualizados, entre outros problemas. Dessa forma, a execução de consultas deve passar a lidar com um mundo probabilístico de acumulação de evidências, em vez de se preocupar em dar respostas exatas a todas as consultas [Abiteboul et al., 2003].

Segurança e Privacidade: Em muitos cenários, proprietários de dados relutam para compartilhá-los sem políticas adequadas de segurança e garantia de privacidade para os mesmos. Pesquisa é necessária para especificar e implementar processos para garantir segurança e privacidade [Halevy, 2003].

6. Conclusões

Integração de dados de múltiplas fontes é um dos problemas mais persistentes na comunidade de pesquisa em banco de dados [Halevy, 2003]. O aparente vasto número de soluções existentes pode, erroneamente, indicar uma área de pesquisa já resolvida. Pelo

contrário, é cada vez mais importante, de forma que ainda não existe uma solução geral que seja adequada ou que se ajuste aos diversos problemas de integração, o que se constata pelo surgimento de novas propostas [Barbosa, 2001].

É importante que os pesquisadores da área de integração de dados passem a trabalhar em conjunto com os das áreas de Inteligência Artificial, Recuperação de Informação, Sistemas Distribuídos e Engenharia de Software [Halevy, 2003], tendo em vista que a relação entre essas três áreas é muito forte, e dessa integração podem surgir soluções importantes para problemas ainda não resolvidos.

Referências

Abiteboul, S., Agrawal, R., Bernstein, P., Carey, M., Ceri, S., Croft, B., DeWitt, D., Franklin, M., Garcia-Molina, H., Galwick, D., Gray, J., Haas, L., Halevy, A., Hellerstein, J., Ioannidis, Y., Kersten, M., Pazzani, M., Lesk, M., Maier, D., Schek, J. N. H., Sellis, T., Silberschatz, A., Stonebraker, M., Snodgrass, R., Ullman, J., Weikum, G., Widom, J., and Zdonik, S. (2003). The lowell database research self assessment.

Arens, Y., Chee, C. Y., Hsu, C.-N., and Knoblock, C. A. (1993). Retrieving and integrating data from multiple information sources. *International Journal of Cooperative Information Systems*, 2(2):127–158.

Barbosa, A. C. P. (2001). *Middleware para Integração de Dados Heterogêneos Baseado em Composição de Frameworks*. PhD thesis, PUC-Rio, Brasil.

Barbosa, A. C. P., Porto, F., and Melo, R. N. (2002). Configurable data integration middleware system. *Journal of the Brasilian Computer Society*, pages 12–19.

Batini, C., Lenzerini, M., and Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv.*, 18(4):323–364.

Beneventano, D., Bergamaschi, S., Guerra, F., and Vincini, M. (2001a). The MOMIS approach to information integration. In *ICEIS (1)*, pages 194–198.

Beneventano, D., Bergamaschi, S., Guerra, F., and Vincini, M. (2001b). Momis: Overview.

Bergamaschi, S., Cabri, G., Guerra, F., Leonardi, L., Vincini, M., and Zambonelli, F. (2001a). The miks architecture overview.

Bergamaschi, S., Cabri, G., Guerra, F., Leonardi, L., Vincini, M., and Zambonelli, F. (2001b). Supporting information integration with autonomous agents. In *Proceedings of the 5th International Workshop on Cooperative Information Agents V*, pages 88–99. Springer-Verlag.

Bergamaschi, S., Castano, S., and Vincini, M. (1999). Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59.

Bertino, E. and Ferrari, E. (2001). Xml and data integration. *IEEE Internet Computing*, 5(6):75–76.

Bouguettaya, A., Benatallah, B., Hendra, L., Ouzzani, M., and Beard, J. (2000). Supporting dynamic interactions among web-based information sources. *IEEE Transactions on Knowledge and Data Engineering*, 12(5):779–801.

Calì, A., Calvanese, D., De Giacomo, G., Lenzerini, M., Naggar, P., and Vernacotola, F. (2003). Ibis: Semantic data integration at work. In *Proc. of the 15th Int. Conf. on Advanced Information Systems Engineering (CAiSE 2003)*, volume 2681 of *Lecture Notes in Computer Science*, pages 79–94. Springer.

Calì, A., Calvanese, D., Giacomo, G. D., and Lenzerini, M. (2002). On the role of integrity constraints in data integration. *Bull. of the IEEE Computer Society Technical Committee on Data Engineering*, 25(3):39–45.

Conrad, S., Eaglestone, B., Hasselbring, W., Roantree, M., Schöhoff, M., Strässler, M., Vermeer, M., and Saltor, F. (1997). Research issues in federated database systems: report of efdbs '97 workshop. *SIGMOD Rec.*, 26(4):54–56.

Cummins, F. A. (2002). *Integração de Sistemas - EAI - Enterprise Application Integration - Arquiteturas para Integração de Sistemas e Aplicações Corporativas*. Editora Campus, Rio de Janeiro - RJ.

de Lima, I. N. (1997). O ambiente web banco de dados: Funcionalidades e arquiteturas de integração. Master's thesis, PUC-Rio, Rio de Janeiro - RJ.

Decker, S., Mitra, P., and Melnik, S. (2000). Framework for the semantic web: An rdf tutorial. *IEEE Internet Computing*, 4(6):68–73.

Elmagarmid, A. K. and Pu, C. (1990). Guest editors' introduction to the special issue on heterogeneous databases. *ACM Comput. Surv.*, 22(3):175–178.

Embley, D. W., Campbell, D. M., Smith, R. D., and Liddle, S. W. (1998). Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 52–59. ACM Press.

Erdmann, M. and Studer, R. (2001). How to structure and access xml documents with ontologies. *Data Knowl. Eng.*, 36(3):317–335.

Florescu, D., Manolescu, I., Kossmann, D., Xhumari, F., and Olteanu, D. (2000). Agora: Living with xml and relational.

Fontes, V., Dutra, M., Porto, F., Schulze, B., and Barbosa, A. (2004). Codims-g: a data and program integration service for the grid. To appear in 2nd International Workshop on Middleware for Grid Computing.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Haas, L. M., Miller, R. J., and et al., B. N. (99). Transforming heterogeneous data with database middleware: Beyond integration.

Hakimpour, F. and Geppert, A. (2001). Resolving semantic heterogeneity in schema integration: an ontology based approach. In *Proceedings of the international conference on Formal Ontology in Information Systems*, volume 2001, pages 297–308, USA.

Halevy, A. Y. (2003). Data integration: A status report. In *Proceedings of 10th Conference on Database Systems for Business Technology and the Web (BTW 2003)*, Germany.

Hsu, C.-N., Chang, C.-H., Siek, H., Lu, J.-J., and Chiou, J.-J. (2003). Reconfigurable web wrapper agents for web information integration. In *Proceedings of IJCAI 2003 Workshop on Information Integration on the Web, IIWeb-03*, pages 15–20.

IBM (2002). "ibm db2 universal database - federated systems guide version 8".

IBM (2003). Db2 information integrator - product overview.

Inmon, W. H. (1997). *Como construir o data warehouse*. Editora Campus, Rio de Janeiro - RJ.

Kreger, H. (2001). Web services conceptual architecture - wsca version 1.0.

Lassila, O. and Swick, R. (1998). Resource description framework (rdf) model and syntax specification.

Linthicum, D. S. (1997). Next generation middleware.

Litwin, W., Mark, L., and Roussopoulos, N. (1990). Interoperability of multiple autonomous databases. *ACM Comput. Surv.*, 22(3):267–293.

Manolescu, I., Florescu, D., Kossmann, D., Xhumari, F., and Olteanu, D. (2000). Agora: Living with xml and relational. In Abbadi, A. E., Brodie, M. L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., and Whang, K.-Y., editors, *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 623–626. Morgan Kaufmann.

Martin, P. and Eklund, P. (1999). Embedding knowledge in web documents. In *Proceeding of the eighth international conference on World Wide Web*, pages 1403–1419. Elsevier North-Holland, Inc.

Mena, E., Kashyap, V., Sheth, A. P., and Illarramendi, A. (1996). OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. In *Conference on Cooperative Information Systems*, pages 14–25.

Miller, R. J., Hernández, M. A., Haas, L. M., Yan, L. L., Ho, C. T. H., Fagin, R., and Popa, L. (2001). The clio project: Managing heterogeneity. 30(1):78–83.

Oracle (2000). Oracle9i integration: The foundation of ebusiness - white paper.

Oracle (2003). Technical comparison of oracle database vs. ibm db2: Focus on information integration.

Pan, A., Raposo, J., Álvarez, M., Montoto, P., Orjales, V., Ardao, J. H. L., Molano, A., and Ángel Via Denodo (2002). The denodo data integration platform.

Parent, C. and Spaccapietra, S. (1998). Issues and approaches of database integration. *Commun. ACM*, 41(5es):166–178.

Park, J. (2001). Schema integration methodology and toolkit for heterogeneous and distributed geographic databases. *Journal of the Korea Industrial Information Systems Society*, 6(3):51–64.

Pitoura, E., Bukhres, O., and Elmagarmid, A. (1995). Object orientation in multidatabase systems. *ACM Comput. Surv.*, 27(2):141–195.

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

Rickett, G. (2001). Pparc e-science opportunities: A call for proposals.

Rodriguez-Martinez, M. and Roussopoulos, N. (2000). Mocha: a self-extensible database middleware system for distributed data sources. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 213–224. ACM Press.

Roth, M. T. and Schwarz, P. M. (1997). Don't scrap it, wrap it! a wrapper architecture for legacy data sources. In *Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 266–275. Morgan Kaufmann Publishers Inc.

Sheth, A. P. and Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surv.*, 22(3):183–236.

Silberschatz, A. and Zdonik, S. (1997). Database systems-breaking out of the box. *SIGMOD Rec.*, 26(3):36–50.

Thomas, G., Thompson, G. R., Chung, C.-W., Barkmeyer, E., Carter, F., Templeton, M., Fox, S., and Hartman, B. (1990). Heterogeneous distributed database systems for production use. *ACM Comput. Surv.*, 22(3):237–266.

Trevisol, G. G. (2004). Codims: Incorporando nova abordagem na comunicação entre seus componentes.

W3C (2004a). Extensible markup language (xml).

W3C (2004b). Tim berners-lee.

Wache, H., ogele, V., Visser, T., Stuckenschmidt, U., Schuster, H., Neumann, G., and ubner, H. (2001). Ontology-based integration of information - a survey of existing approaches.

Wiederhold, G. (1998). Value-added middleware: Mediators.

Wiederhold, G. and Genesereth, M. (1997). The conceptual basis for mediation services. *IEEE Expert*, 12(5):38–47.

Xhumari, F., Amzal, M., and Simon, E. (2000). Le select: a middleware system for publishing autonomous and heterogeneous information sources. Technical report, INRIA, French.

Yuan, X. (1998). Interoperability of heterogeneous geographic information processing environment for internet gis. *Journal of Wuhan Technical University of Surveying and Mapping*.

Özsu, M. T. and Valduriez, P. (2001). *Princípios de Sistemas de Banco de Dados Distribuídos*. Editora Campus, Rio de Janeiro - RJ.