Taxonomia para projetos de integração de fontes de dados baseados em ontologias

Taxonomy for data source ontology-based integration projects

Mauricio B. Almeida[1]
Marcello P. Bax[2]

Resumo

Nos últimos anos, vários trabalhos na literatura tem abordado os problemas de integração de fontes de dados em ambientes abertos e heterogêneos. A importância desse problema deriva do fato de que, com um número crescente de fontes de dados disponíveis, torna-se cada vez mais difícil a seleção, aquisição e combinação de dados. Dentre as propostas da literatura, um grande número utiliza ontologias como ferramenta de integração. Essas propostas apresentam características muito distintas o que torna difícil a comparação direta entre os projetos. Nesse artigo, propõe-se uma taxonomia para caracterizar projetos de integração de fontes de dados baseados em ontologias. Pesquisa-se as principais iniciativas descritas na literatura e realiza-se uma análise qualitativa. Espera-se que esse trabalho seja um estímulo para estudos mais abrangentes das abordagens de integração de fontes de dados em outras áreas do conhecimento.

Palavras chave: ontologia, integração, inter-operabilidade

Abstract

In the past years, several works in the literature have addressed problems of data sources integration in open and heterogeneous enviromments. The importance of this problem derives from the fact that, with the increasing number of the available data sources, it is more and more dificult to make data selection, data acquisition and data combination. There are many approaches in the literature which use ontologies like integration tools. These approaches present distinct features and it is dificult to make a direct comparison between the projects. In this paper, we propose a taxonomy to characterize ontology-based integration projects. We survey briefly the most important initiatives described in the literature and make a qualitative analysis. Hopefully, this work will estimulate other more comprehensive studies about the approaches of data sources integration in other knowledge fields.

Keywords: ontology, integration, interoperability

1 Introdução

A competitividade das organizações no ambiente de negócios atual, depende da qualidade do acesso às informações que a empresa manipula rotineiramente. A tarefa de proporcionar acesso essas informações tem se mostrado árdua. A Internet, *intranets*, redes de alta velocidade e infra-estruturas de computação distribuída continuam a ganhar em popularidade como meios eficientes de comunicação. Em tais ambientes, caracterizados por fontes de dados altamente distribuídas, o acesso à informação relevante torna-se cada vez mais complexo.

Um número cada vez maior de fontes de dados está disponível *on-line* possibilitando acesso

[1] Mestre em Ciência da Informação pela ECI/UFMG e professor assistente da PUC MINAS. mba@pucminas.br .
[2] Doutor em Ciência da Computação e professor adjunto da ECI-UFMG. bax@ufmg.br .

mais fácil e novas combinações de dados. Essas fontes nem sempre podem ser facilmente integradas em função de diversos tipos de heterogeneidade que podem ocorrer (OUKSEL; SHETH, 1999; WATSON, 2000; WEIHAI, 2002; BALDONADO; COUSINS, 1997; SHETH, 1998; OMELAYENKO, 2001).

O simples acúmulo de um número cada vez maior de fontes de dados disponíveis *on-line* conduz à "sobrecarga de informações" (WIEDERHOLD, 1994). O usuário não consegue mais extrair informação de um conjunto de dados recuperados. Dessa forma, um dos objetivos da integração no ambiente informacional é buscar o aumento do valor da informação no momento em que dados de várias fontes são acessados, relacionados e combinados.

Existem abordagens que tratam problemas de integração entre fontes de dados utilizando estruturas de organização do conhecimento conhecidas como ontologias. Em conjunto com as ontologias, utilizam-se a lógica como forma de representação do conhecimento (que pode ser implementada em computadores) e linguagens de programação desenvolvidas pela computação para a construção de sistemas.

Ontologias são utilizadas hoje em diversas áreas para organizar conhecimento (BATEMAN, 1996; BORGO et al., 1997; AGUADO et al., 1998; DOMINGUE et al., 1998; HASMAN et al., 1999; SHUM; MOTTA; DOMINGUE, 2000; LEGER et al., 2000; KALFOGLOU, 2001; VÁZQUEZ; VALERA; BELLIDO, 2001; GANDON, 2001; MARTIN; EKLUND, 2001; ALEXAKI et al., 2002). Os projetos para integração de fontes de dados baseados em ontologias apresentam características diversas. Acredita-se não ter sido produzida ainda uma análise abrangente das abordagens dos projetos de integração, em função da dificuldade em estabelecer um quadro genérico para comparação. Esse artigo é uma primeira tentativa nessa direção.

Este artigo está organizado conforme segue: a Seção 2 introduz conceitos básicos e características das ontologias; a Seção 3 apresenta a taxonomia para caracterização de projetos de integração que utilizam ontologias; na Seção 4, apresenta-se uma visão geral dos projetos de integração baseadas em ontologias citados na literatura; a Seção 5, apresenta uma análise qualitativa

desses projetos. Finalmente, a seção 6 apresenta as conclusões e aponta direções para trabalhos futuros.

2 Ontologias

Essa seção apresenta uma breve definição para ontologia e suas possíveis categorias, destacando o significado tradicional do termo (utilizado na filosofia) e o significado do termo no contexto desse artigo. Apresenta-se também a idéia de como as ontologias podem ser aplicadas aos processos de integração de fontes de dados.

2.1 Definição de ontologia

Historicamente o termo ontologia tem origem no grego *ontos*, ser e *logos*, palavra. É um termo introduzido na filosofia[3] com o objetivo de distinguir o estudo do ser humano como tal, do estudo de outros seres das ciências naturais. A origem é a palavra aristotélica "categoria", que pode ser usada para classificar e caracterizar alguma coisa.

O termo ontologia tem um sentido especial nos projetos que são aqui estudados. Mesmo considerando-se apenas o domínio da computação, de onde têm se originado muitos trabalhos sobre o tema, são diversos os conceitos apresentados para o termo e existem muitas contradições (GUARINO; GIARETTA, 1995), (GUARINO, 1996), (GRUBER, 1996), (ALBERTAZZI, 1996), (USCHOLD; GRUNINGER, 1996), (NECHES et al., 1991), (CHANDRASEKARAN, JOHNSON, BENJAMINS, 1999).

Borst (1997, p. 12) apresenta uma definição simples e completa, a qual será adotado nesse artigo: "Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada". Nessa definição, "formal" significa legível para computadores; "especificação explícita" diz respeito a conceitos, propriedades, relações, funções, restrições, axiomas que são explicitamente definidos; "compartilhado" quer dizer conhecimento consensual; e, "conceitualização" diz respeito a um modelo abstrato de algum fenômeno do mundo real.

---

[3] A definição do dicionário Oxford de Filosofia é "[...] o termo derivado da palavra grega que significa 'ser' [,...]". O Dicionário Aurélio traz: "Ciência do ser em geral."

As ontologias podem ser classificadas de acordo com o grau de formalidade de seu vocabulário (USCHOLD; GRUNINGER, 1996) em relação a estrutura e ao assunto da conceitualização (VAN-HEIJIST, SCHREIBER; WIELINGA, 1997), em relação a sua função (MIZOGUCHI, VANWELKENUYSEN; IKEDA, 1995), (HAAV; LUBI, 2001), em relação a sua aplicação (JASPER; USCHOLD, 1999).

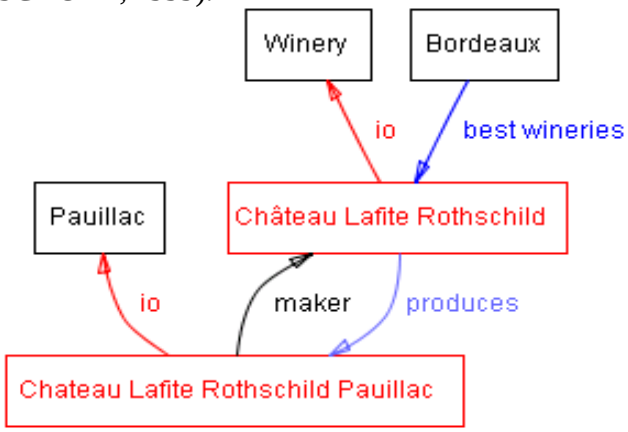

Figura 1 – Fragmento de uma ontologia para tipos de vinhos

Fonte: NOY; GUINNESS, 2001.

2.2 Ontologia como ferramenta de integração

Em relação ao uso de uma ontologia em um sistema, identificam-se possíveis categorias: ontologias de autoria neutra (enfatizam a reutilização de informações), ontologias como especificação (enfatizam documentação e manutenção) e ontologias de acesso comum a informação (enfatizam o acesso a informação) (JASPER; USCHOLD, 1999). A última categoria se aplica quando a informação desejada é expressa em um vocabulário inacessível e a ontologia possibilita o seu entendimento, proporcionando conhecimento compartilhado dos termos ou inter-relacionando grupos de termos.

O acesso comum a informações a partir da ontologia está relacionado a sua aplicação como ferramenta de integração. A idéia de que técnicas de organização, como o uso de ontologias, podem auxiliar na integração entre fontes de dados têm sido bastante discutida. Como as ontologias possibilitam uma compreensão comum e compartilhada de um domínio do conhecimento, em que

deve haver comunicação entre pessoas e sistemas, elas têm papel importante no intercâmbio de informações, pois proporcionam uma estrutura semântica às fontes de dados. Dessa forma, é possível a comunicação entre os agentes envolvidos nos processos (computadores ou pessoas), ao serem reduzidas diferenças conceituais ou terminologias.

3 Taxonomia para projetos de integração que usam ontologias

Esta seção apresenta uma taxonomia que reúne os projetos estudados em grupos, baseando-se no papel que as ontologias desempenham na descrição do conteúdo das fontes de dados durante o processo de integração. Esse critério leva aos seguintes grupos: ontologia global (usa-se uma ontologia única), multi-ontologias (usam-se diversas ontologias) e ontologias combinadas (usa-se ontologias globais associadas a outras ontologias).

Apesar da utilidade dessa taxonomia para propósitos didáticos, deve se considerar a possibilidade de diversas outras. Na verdade, podem existir casos em que um sistema pertence a mais de um grupo. A divisão em grupos que compõem a taxonomia apresentada na Figura 2 se baseia nos estudos de (WACHE, 2001) e (STUCKENCHMIDT et al., 2001):

Taxonomia	Papel da ontologia	Projetos estudados
Ontologia global	Todas as fontes de dados são relacionadas a única ontologia global desenvolvida independentemente das outras fontes e de suas ontologias. A ontologia global faz o mapeamento inter-ontologias e representa um vocabulário compartilhado para especificação da semântica.	TSIMMIS (CHAWATHE, 1994); SIMS (ARENS, HSU e KNOBLOCK, 1996); KRAFT (PREECE e HUI, 2001): BUSTER (STUCKENSCHMID E WACHE, 2000)
Multi-ontologias:	Cada fonte de informação é representada por sua própria ontologia, não sendo necessária uma ontologia global. As ontologias das fontes de dados podem ser desenvolvidas sem preocupação com as outras fontes ou com suas respectivas ontologias. A ausência de um vocabulário comum torna difícil a comparação entre diferentes ontologias. Assim é necessário um formalismo de representação definindo uma relação "inter-ontologias".	OBSERVER (MENA et al., 1996); ONTOBROKER (FENSEL et al., 1998); SHOE (HEFLIN e HENDLER, 2000).
Ontologias combinadas	A semântica de cada fonte é descrita por sua própria ontologia. Para tornar as ontologias locais comparáveis, elas são construídas a partir de um vocabulário global compartilhado, que pode inclusive ser outra ontologia.	COIN (BRESSAN, 1997),; INFOSLEUTH (NODINE, BOHRER e NGU, 1999): PICSEL(GOASDOUÉ; LATTES; ROUSSET, 1998); DWQ (CALVANESE, 1998);

Figura 2 – Tanoxomia proposta, características básicas de cada grupo e projetos estudados

Os projetos possuem diferentes características, princípios de funcionamento e diferentes papeis para a ontologia, assunto que é apresentado na próxima seção.

4 Visão geral dos projetos de integração pesquisadas

Nessa seção, apresenta-se uma visão geral dos projetos de integração citados, baseados em ontologias. A lista de projetos aqui apresentada não tem a pretensão de ser completa e apesar da preocupação em cobrir os projetos mais representativas descritas recentemente na literatura, este estudo não é exaustivo. A apresentação a seguir esta ordenada de acordo com a taxonomia descrita na Seção 3.

4.1 Uso de ontologia global

4.1.1 TSIMMIS - *Stanford/IBM Manager of Multiple Information Sources*

O *TSIMMIS* é um projeto pioneiro de integração concebido pela comunidade de banco de dados. Desenvolve ferramentas para a integração de fontes de dados estruturados ou não.

Para cada fonte disponível, o sistema aloca um *wrapper*[4] que converte objetos representativos dos dados em um modelo comum (CHAWATHE et al., 1994). Este modelo auto-descritivo (os objetos têm rótulos que descrevem seu significado) é chamado *Objetc Exchange Model (OEM)*. O *wrapper* converte as consultas executadas no formato do modelo comum, em pedidos que a fonte pode executar. Os dados retornam da fonte convertidos em dados do modelo comum.

Mediadores[5] fazem correspondência com o modelo *OEM* (MENA et al., 1996). Um mediador processa respostas antes de encaminhá-las ao usuário, convertendo dados para o formato comum e eliminando redundâncias. O mediador trabalha de forma independente das fontes que vai usar. *Wrappers* e mediadores recebem como entrada consultas em uma linguagem específica (OEM-*Query Language*) e retornam objetos *OEM*. Os usuários finais podem acessar a informação escrevendo aplicações que recebem objetos *OEM*.

4.1.2 SIMS - *Services and Information Management for decision Systems*

O *SIMS* é um mediador que acessa e integra fontes de dados. As consultas são expressas em

---

4 Elemento da arquitetura faz a tradução entre a fonte e o modelo utilizado no sistema. Um *wrapper* deve ser escrito para cada tipo de fonte, mas já existem abordagens semi-automáticas para sua geração.

5 Elemento da arquitetura que recebe informações do *wrapper* e as transmite para outras partes do sistema, após algum processamento.

uma linguagem comum, independente das possíveis linguagens de consulta existentes e da localização das fontes (ARENS; HSU; KNOBLOCK, 1996).

O mediador contém um modelo do domínio (base de conhecimento hierárquica e terminológica) e um modelo de todas as fontes de informação. A modelagem descreve o relacionamento entre classes, subclasses e superclasses, a função de cada classe, o seu conteúdo e a integração desses modelos com o modelo do domínio.

A fonte de dados é selecionada para atender a uma consulta expressa em termos do modelo do domínio. O sistema aplica operadores para transformar conceitos do domínio em outros, que podem ser recuperados diretamente da fonte de dados. Uma vez que o sistema reformulou as consultas para usar termos dos modelos das fontes, gera-se um plano de consultas para recuperação e processamento dos dados, o qual especifica quais operações e em que ordem devem ser executadas. A transformação de uma consulta em outra é feita através de inferências lógicas usando-se abstrações que descrevem os bancos de dados com um conjunto de formulas de lógica de 1ª ordem[6].

4.1.3 KRAFT - *Knowledge Reuse And Fusion/Transformation*

O *KRAFT* adota a abordagem de *fusão de conhecimento*, processamento que associa e combina dados de várias fontes. Caracteriza-se pelo refinamento das estimativas e avaliação dinâmica da necessidade de fontes adicionais.

As fontes de dados individuais são atribuídas a seus esquemas locais e ao esquema de integração, para que os dados possam ser combinados (PREECE; HUI, 1999). Para essa combinação usa-se o conhecimento associado ao contexto. Este tipo de conhecimento pode ser expresso na forma de *restrições*[7].

No KRAFT as instâncias de dados e respectivas restrições são atribuídas a uma representação comum permitindo a fusão. É utilizada uma linguagem comum para representação das instâncias e restrições e um grupo comum de definições de terminologia do domínio de conhecimento (uma

---

6 Linguagem que descreve verdades lógicas por fórmulas matemáticas e utiliza “conectivos”.

7 Uma restrição pode ser um atributo de um dado que o restringe a uma faixa de valores

ontologia compartilhada). O conhecimento presente nos recursos individuais precisa ser transformado na linguagem comum, nos termos da ontologia, antes que possa ser combinado.

Os componentes de processamento de conhecimento da arquitetura do *KRAFT* são *agentes* de *softwares*, de três tipos: os *wrappers* (ligação com fontes de dados externas), mediadores (processamento interno do conhecimento obtido por outros agentes e facilitadores) e facilitadores (proporcionam a comunicação entre agentes).

4.1.4 BUSTER - *Bremen University Semantic Translator*

No *BUSTER*, o vocabulário compartilhado é uma ontologia global e uma ontologia que representa a fonte é um refinamento (parcial) da ontologia geral que a restringe em uma faixa de valores de alguns atributos. Como as ontologias da fonte usam apenas o vocabulário da ontologia geral, as duas permanecem comparáveis.

O *BUSTER* parte do principio que é necessário analisar e transformar o contexto do conhecimento, para que se possa alcançar integração em nível semântico. Procura preservar o significado de atributos únicos no banco de dados, visto que terminologias diferentes podem ser usadas (STUCKENSCHMIDT; WACHE, 2000).

Regras de integração orientam a correlação das fontes e se obtém uma visão integrada fornecida pelo mediador. As regras se baseiam na descrição de objetos de cada fonte. Assim, um objeto, tabela de banco relacional ou um número, será “encapsulado” por um objeto padrão. Este objeto é um predicado, conforme definido na lógica, com campos que contém variáveis e descrevem instâncias encontradas no banco de dados.

4.2 Uso de multi-ontologias

4.2.1 OBSERVER - *Ontology Based System Enhanced with Relationships for Vocabulary heterogeneity Resolution*

O conteúdo das fontes de dados é descrito por conceitos ontológicos que representam um domínio. O problema na integração das fontes é compartilhar o vocabulário, ou seja, lidar com

diferentes termos ou conceitos para descrever informações similares (MENA et al, 1996). O *OBSERVER* permite a representação de relacionamentos "inter-ontologias" e correlaciona os termos entre elas. As consultas do usuário são reescritas usando esses relacionamentos, o que resulta em traduções entre as ontologias.

As ontologias são diferentes pois são desenvolvidas de forma independente por organizações diferentes e porque vocabulários diferentes atendem melhor às necessidades dos usuários. O *OBSERVER* usa múltiplas ontologias que estão ligadas por relacionamentos inter-ontologias e organizadas em grupos que correspondem as áreas do conhecimento. Como alguns grupos são mais gerais que outros, podem ser organizadas hierarquias.

As ontologias são descritas por um sistema de lógica descritiva, tornando-se reutilizáveis após a representação. Mesmo tendo sido criadas com diferentes linguagens de representação, a lógica descritiva mantém a semântica original. Estas ontologias são usadas para descrever repositórios de dados.

4.2.2 ONTOBROKER

O *ONTOBROKER* é uma arquitetura que implementa ferramentas necessárias ao uso de ontologias para realizar consultas na Internet. A arquitetura é composta por três elementos principais: uma interface (consulta de usuários), uma máquina de inferência (usada para se obter respostas) e um agente inteligente (utilizado para coletar dados da *Web*) (FENSEL et al., 1998).

O *ONTOBROKER* utiliza a abordagem de "meta-anotações" que adicionam informações semânticas às fontes de informação. Em geral, meta-anotações consistem de marcações especiais em linguagens conhecidas da Internet (extensões), ou do desenvolvimento de uma linguagem específica. Essa abordagem tem sido muito utilizada para integração fontes de dados na Internet, onde anotação é uma forma natural de proporcionar semântica.

Utiliza uma linguagem de representação para a criação de ontologias, da qual, um subgrupo é utilizado para elaborar consultas. Possui ainda uma "linguagem de anotação" para permitir ao fornecedor "marcar" semanticamente documentos da Internet com informações ontológicas.

4.2.3 SHOE - *Simple HTML Ontology Extensions*

O *SHOE* utiliza meta-anotações que adicionam semântica às fontes, de forma similar ao *Ontobroker* (Seção 4.2.2) A idéia é possibilitar que o autor de um documento possa inserir metadados diretamente na página, melhorando a recuperação da informação. As marcações adicionam conhecimento ao contexto pois são "marcações semânticas", definidas em um grupo de atributos e relacionamentos (ou seja, uma ontologia).

O *SHOE* consiste de extensões à linguagem *HTML-Hypertext MarkUp Language* que permitem que autores de páginas da *Web* façam anotações, proporcionando correlações semânticas (LUKE et al., 1997). Estas anotações são expressas em conhecimento ontológico e isso possibilita a execução eficiente de consultas, via agentes inteligentes.

O *SHOE* proporciona ainda a definição de ontologias usando *HTML*, a criação de novas ontologias que podem complementar ontologias existentes, declaração de entidades e relacionamentos entre entidades e classificação entidades em um esquema do tipo "*is-a*".

4.3 Uso de ontologias combinadas

4.3.1 COIN - *Context Interchange Project*

O *COIN* integra de fontes de dados através de uma arquitetura de mediadores. Um mediador proporciona consultas às fontes e tem a capacidade de resolver conflitos semânticos. *Wrappers* fazem a ligação entre as fontes de dados e o mediador (BRESSAN, 1997).

Os componentes do *COIN* executam três grupos diferentes de processos: processo-cliente (proporcionam interação entre os pedidos ao banco de dados e o mediador); processo-servidor (*wrappers* permitem consultas a documentos semi-estruturados na Web): processo-mediadores (reescreve consultas do usuário em uma consulta mediada, envia sub-consultas aos processos servidores, opera resultados intermediários e retorna respostas finais).

O *COIN* adota uma representação "*frame-based*"[8] aliada a uma sintaxe definida pelo próprio

[8] Baseada em "frames", estruturas que contém variáveis pertencentes a um escopo.

sistema.

4.3.2 INFOSLEUTH - *Intelligent Search Management via Semantic Agents*

O *INFOSLEUTH* utiliza uma rede de agentes para coleta e análise de dados em um sistema global (por exemplo, a Internet), recuperando e processando dados. Integra agentes, ontologias e computação distribuída, para mediação de dados em um ambiente dinâmico (MENA et al, 1996)

Uma ontologia especial é utilizada para comunicação dos agentes e outras ontologias são usadas para a captura de dados. A consulta é expressa em conceitos ontológicos interpretados por agentes. A comunicação entre os agentes retornam uma resposta ao usuário (HWANG, 1999). Os usuários especificam consultas nas ontologias via uma interface. A linguagem de representação do conhecimento KIF- *Knowledge Interchange Format* e a linguagem de consulta de banco de dados SQL-*Structured Query Language* são utilizadas internamente para representar consultas em ontologias específicas.

A arquitetura do *INFOSLEUTH* é baseada nos seguintes agentes: do usuário (utiliza ontologias para a formulação de consultas); da ontologia (disponibiliza o conhecimento das ontologias); de intermediação (armazena mensagens de agentes sobre suas capacidades e direciona pedidos); de recursos (correlaciona a ontologia comum com o esquema do banco de dados e sua linguagem nativa); de análise de dados (coleta informação); agente monitor (fornece interface para exibir interações) (BAYARDO et al., 1997).

4.3.3 PICSEL

O *PICSEL* consiste de um mediador composto de duas partes: por uma *máquina de consultas* e por *bases de conhecimento*[9] específicas, as quais contêm modelos do domínio e descrições do conteúdo das fontes de informação acessíveis (GOASDOUÉ; REYNAUD, 1999). O modelo de domínio contém todo o vocabulário para responder as consultas. O motor de consultas acessa fontes para responder às consultas.

---

[9] Conhecimento expresso com a utilização de alguma linguagem formal de representação do conhecimento. Uma base de conhecimento faz parte de um sistema baseado em conhecimento.

O conteúdo das fontes de informação é representado no mesmo formalismo lógico das consultas e do domínio. *Wrappers* trabalham de forma especializados para cada modelo de dados. Quando uma fonte é um banco de dados relacional, *wrappers* traduzem a consulta em termos das relações das fontes para uma forma relacional.

O *PICSEL* usa uma linguagem, que contém grupos de regras e grupos de declarações em lógica descritiva que são definições sobre conceitos e papéis no domínio, para representar o conteúdo das fontes de dados disponíveis

Durante a integração as correlações são mapeadas do esquema global para o esquema da fonte local. Uma sub-consulta está correta se fornece uma parte das respostas solicitadas (isto é, as sub-consultas devem estar contidas na consulta global). Como uma ontologia possui uma especificação completa da conceitualização (GRUBER, 1993), as correlações podem ser validadas pela ontologia. Os conceitos da ontologia correspondentes às sub-consultas locais estão contidos nos conceitos da ontologia relacionados à consulta global.

4.3.4 DWQ - *Datawarehouse Quality*

No *DWQ*, além da função de análise do conteúdo, a ontologia tem a tarefa de descrever a integração. O processo de integração utiliza representação em lógica descritiva. Assume-se que cada fonte é uma coleção de tabelas relacionais e cada uma delas é descrita em termos de sua ontologia. Uma consulta global e sua decomposição em sub-consultas é relacionada a conceitos ontológicos. As sub-consultas estão corretas, isto é, estão contidas na consulta global, se seus conceitos ontológicos estão contidos nos conceitos da ontologia global.

O projeto *DWQ* proporciona a base semântica para o projeto de grandes bancos de dados utilizando modelos complexos e estruturas semanticamente ricas, de maneira sistemática, facilitando o projeto, operação e evolução dos bancos de dados (JARKE; VASSILIOU, 1997).

A arquitetura do *DWQ* abrange o projeto, a configuração, a operação, manutenção de *datawarehouses*. Os elementos da arquitetura são: fontes (repositório que pode ser utilizado como fonte de dados); *wrappers* (descrevem as fontes no formato da ontologia genérica)*;* bancos destino

(*datawarehouses)*; banco de metadados (repositório para informação sobre outros componentes); agentes administrativos; clientes (apresentação dos dados).

5 Análise qualitativa dos projetos de integração

Nessa seção, analisa-se como os projetos estudados são caracterizadas em relação a ontologia e ao processo de integração. Estudam-se as seguintes características: representação da ontologia, mapeamento inter-ontologias, conexão com a fonte de informação, modelo de dados, arquitetura e princípios de funcionamento.

5.1 Representação da ontologia

A representação do conhecimento de ontologias utilizadas no processo de integração, pode ser feita por linguagens específicas.A importância das características da linguagem de representação reside em conhecer o poder de expressividade da linguagem e assim, determinar qual é mais adequada ao contexto. Estudar ou avaliar as características de cada linguagem está além do escopo desse trabalho.

As linguagens dominantes para esse fim são variantes da lógica descritiva. Exemplos de lógicas descritivas "puras" são a CLASSIC (lógica descritiva), a GRAIL (RECTOR et al., 1997) e a OIL (FENSEL et al., 2000). Exemplos de extensões das lógicas descritivas são a CARIN (LEVY; ROUSSET, 1996) e DLR (CALVANESE, 1998). Outro grupo de linguagens para a representação de ontologias são as linguagens conhecidas como *frame-based* , dentre as quais tem-se *F-logic* (KIFER; LAUSEN; WU, 1990) e *Ontolingua* (CHAUDHRI et al., 1998).

Dentre os projetos que executam a integração através de uma única ontologia global, o SIMS utiliza a LOOM (BRILL, 1993), o KRAFT utiliza a *CoLan - Constraint Language* (linguagem baseada em restrições) e o BUSTER utiliza a OIL. Dentre os projetos que executam a integração através várias ontologias, o OBSERVER utiliza a CLASSIC, o ONTOBROKER utiliza a *F-Logic* e o SHOE utiliza uma linguagem de marcação, extensão do HTML, para especificação de metadados. Dentre os projetos que executam a integração através várias ontologias combinadas, o COIN utiliza a

F-logic, o INFOSLEUTH utiliza a KIF, o PICSEL utiliza a CARIN, o DWQ utiliza a DLR.

5.2 Mapeamento inter-ontologias

Usa-se o termo mapeamento para expressar a conexão de uma ontologia com as outras partes de um sistema (WACHE, 2002). No caso do mapeamento inter-ontologias, o sistema de integração utiliza mais de uma ontologia para descrever a fonte de informação.

Dentre os projetos que executam a integração através de uma única ontologia global, o KRAFT executa a tradução entre as ontologias através de um agente mediador especial customizável, em um processo que não preserva a semântica, a qual pode ser alterada pelo usuário. No BUSTER, procura-se resolver o problema anterior tentando identificar a correspondências semânticas entre conceitos de diferentes ontologias.

Dentre os projetos que executam a integração através várias ontologias, o OBSERVER, utiliza um modelo de lógica descritiva utilizando relacionamentos inter-ontologias baseados em conceitos lingüísticos (por exemplo, sinônimos). Apesar de semelhantes às construções da lógica descritiva, não tem uma semântica formal, o que sugere o uso de heurísticas nos algoritmos.

Dentre os projetos que executam a integração através várias ontologias combinadas, o DWQ, relaciona a ontologia ao formalismo de uma ontologia de nível mais alto, evitando assim conflitos e ambigüidades pela perda de semântica. Essa abordagem estabelece ligações entre os conceitos de várias ontologias, mas não estabelece uma correspondência direta.

5.3 Conexão com a fonte de informação

Dentre os projetos que executam a integração através de uma única ontologia global, o TSIMIMS e o SIMS, efetuam a conexão da ontologia com a fonte produzindo uma cópia da estrutura do banco de dados. A integração é executada sobre essa cópia do modelo e pode ser retornar aos dados originais facilmente. No BUSTER, usa-se a ontologia para definir termos do banco de dados ou de seu esquema. Estas definições (conjunto de regras) não correspondem a estrutura do banco de dados, mas apenas ligam os dados aos termos que os definem. No KRAFT é construído um modelo

lógico que contém cópia da estrutura e definições dos conceitos.

Dentre os projetos que executam a integração através várias ontologias, o OBSERVER adota uma abordagem intermediária em relação a utilizada no KRAFT. O ONTOBROKER e o SHOE utilizam meta-anotações para adicionar informações às fontes de dados.

Dentre os projetos que executam a integração através várias ontologias combinadas, o PICSEL e o DWQ também utilizam abordagem similar à usada no KRAFT.

5.4 Arquitetura e princípios de funcionamento

A maioria dos projetos utiliza uma arquitetura de *wrappers* e mediadores (TSIMMIS, SIMS, COIN, KRAFT, PICSEL). Existem projetos que utilizam agentes de software (INFOSLEUTH e KRAFT), projetos que utilizam princípios léxicos aliados a um mediador (OBSERVER), projetos que se baseiam em meta-anotações (ONTOBROKER e SHOE) e projetos que utilizam regras de interação (BUSTER).

5.5 Quadros sinóticos

Para maior clareza, apresentam-se dois quadros sinóticos, um com a classificação taxonômica proposta e outro com a análise qualitativa. O quadro sinótico da figura 3 apresenta as mesmas informações da Figura 2, mas seu formato possibilita visualização mais rápida do projeto e de seu respectivo grupo taxonômico.

Taxo-nomia	**Projeto**										
	TSI MMIS	SIMS	COIN	OBSER VER	INFO SLEUTH	KRAFT	PICSEL	DWQ*	ONTO BROKER	SHOE	BUSTER
Ontologia global	●	●				●					●
Multi-ontologias				●					●	●	
Ontologias combinadas			●		●		●	●			

Figura 3 – papel da ontologia em cada projeto de integração

Taxonomia	Projetos	Análise qualitativa			
		Representação da ontologia	Mapeamento inter- ontologias	Conexão com a fonte	Arquitetura e funcionamento
Ontologia global	TSIMMIS	-	Não	Cópia da estrutura do banco de dados	Arquitetura de wrappers e mediadores
	SIMS	Lógica descritiva *LOOM*	Não	Cópia da estrutura do banco de dados	Arquitetura de wrappers e mediadores
	KRAFT	*Colan*	Sim	Modelo - cópia da estrutura e conceitos	Arquitetura de wrappers e mediadores(agentes)
	BUSTER	Lógica descritiva OIL	Sim	Conjunto de regras	Usa regras de interação
Multi-ontologias	OBSERVER	lógica descritiva *CLASSIC*	Sim	Modelo - cópia da estrutura e conceitos	Mediador usa princípios léxicos
	ONTOBROKER	*Frame-baased F-logic*	Não	Uso de meta-anotações	Meta-anotações enriquecem o conteúdo da fonte
	SHOE	Extensão HTML	Não	Uso de meta-anotações	Meta-anotações enriquecem o conteúdo da fonte
Ontologias combinadas	COIN	*Frame-baased F-logic*	Não	-	Arquitetura de wrappers e mediadores
	INFOSLEUTH	Linguagem KIF	Não	-	Arquitetura de agentes
	PICSEL	lógica descritiva CARIN	Não	Modelo - cópia da estrutura e conceitos	Arquitetura de wrappers e mediadores
	DWQ	lógica descritiva Extensão DLR	Sim	Modelo - cópia da estrutura e conceitos	-

Figura 4 – resumo da análise qualitativa de cada projeto

6 Conclusões e trabalhos futuros

Apresentou-se uma pesquisa de projetos para a integração de fontes de dados baseados em ontologias. Introduziu-se uma taxonomia para classificação dos projetos estudados de acordo com o papel da ontologia na descrição do conteúdo. Em alguns casos, como no DWQ, a ontologia pode ser implementada em qualquer um dos três papeis citados.

Analisou-se qualitativamente os projetos estudados, examinando-se algumas de suas características importantes em projetos de integração como: representação do conhecimento, mapeamento inter-ontologias, conexão com a fonte de informação e arquitetura e princípios de funcionamento. Nas Figuras 3 e 4 apresentaram-se quadros sinóticos que resumem as principais características dos projetos estudados.

As características apresentadas são diversas e o artigo procurou realizar uma primeira abordagem no sentido de proporcionar a comparação entre os projetos. Estudos mais abrangentes poderão dar seguimento a essa iniciativa.

Conclui-se a partir do estudo das características dos projetos que provavelmente o esforço de

implementação deva ser menor em sistema que utilizam uma ontologias global. Entretanto, nos sistema que utilizam a abordagem multi-ontologias, parece ser mais simples adicionar novas fontes de dados, sem necessidade de alterações em uma ontologia única. A comparação entre as ontologias parece ser difícil nos projetos multi-ontologias, em função a não existência de um vocabulário compartilhado, o que não parece acontecer nos projetos de ontologias combinadas, onde se preserva a semântica através do vocabulário compartilhado.

O mapeamento semântico proporcionado pela ontologias poderá auxiliar na integração de fontes de dados heterogêneas, hoje muito comuns em ambientes abertos e distribuídos como a Internet e as Intranets. Dessa forma, espera-se conseguir melhorias nos processos de recuperação de informação, proporcionando às empresas maior qualidade no acesso e manipulação de dados.

Em trabalhos futuros, espera-se abordar princípios de engenharia para a construção de ontologias, visto que não se trata de uma tarefa trivial. Qualquer automatização no processo de construir ontologias poderá ser de utilidade para que possam ser utilizadas efetivamente em projetos de integração em ambientes heterogêneos. Além disso, a espera-se abordar em trabalhos futuros o uso dos princípios de funcionamento aqui estudados, no âmbito da Web Semântica Corporativa (GANDON, 2001). Essa abordagem poderá propiciar ao ambiente corporativo, as vantagens obtidas com os projetos de integração utilizados em ambientes abertos.

7 Referências bibliográficas

AGUADO, G. et al. Ontogeneration: Reusing domain and linguistic ontologies for Spanish text generation? In: 13th EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, ECAI'98. *Papers Accepted to the Workshop on Applications of Ontologies and Problem solving Methods*. Brighton, England, p. 23-28, august 1998.

ALBERTAZZI, Liliana. Formal and material ontology. In: POLI, Roberto; SIMONS, Peter (Ed.). *Formal Ontology*. Dordrecht: Kluwer, 1996. p. 199-232.

ALEXAKI, S. et al. Managing RDF Metadata for Community Webs. In: 2nd INTERNATIONAL WORKSHOP ON THE WORLD WIDE WEB AND CONCEPTUAL MODELING. p. 140-151, 2000. Disponível em: <http://139.91.183.30:9090/RDF/ publications/wcm2000.PDF>. Acesso em: 11 out. 2002.

ARENS, Y.; HSU, C. N.; KNOBLOCK, C. A. Query processing in the SIMS information mediator. In: TATE, Austin. *Advanced Planning Technology*: technological achievements of the ARPA/Rome Laboratory planning inititive. Menlo Park, AAAI Press, 1996. 290 p.

BALDONADO, M.; COUSINS, S. Addressing heterogeneity in the networked information environment. *New Review of Information Networking*, v. 2, p. 83-102, 1996.

BATEMAN, J. A. (1996). *Using text structure and text planning to guide text summarization*. Disponível em: <http://www.ik.fhhannover.de/ik/projekte/Dagstuhl/ Abstract/Abstracts/Bateman/Bateman.html>. Acesso em: 25 maio 2002.

BAYARDO, R. J. et al. INFOSLEUTH: agent-based semantic integration of information in open and dynamic environments. In: PROCEEDINGS OF THE 1997 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT. ACM SIGMOD Record, v. 26, Issue 2, june 1997.

BORGO, S. et al. Using a Large Linguistic Ontology for Internet-Based Retrieval of Object-Oriented Components. In: PROC. OF 9th INT. CONF. ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING (SEKE 97). Madrid, Spain, 1997.

BORST, W.N. (1997). *Construction of Engineering Ontologies*. Phd Thesis. Disponível em: <http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf>. Acesso em: 21 abr. 2002.

BRESSAN, S. Semantic Integration of Disparate Information Sources over the Internet using Constraint Propagation. In: WORKSHOP ON CONSTRAINT REASONING ON THE INTERNET AT CP-97. Schloss Hagenberg, Austria, October 29 - November 1 1997.

CALVANESE, D. et al. Description logic framework for information integration. In: PROCEEDINGS OF THE SIXTH INTERNATIONAL CONFERENCE ON PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING (KR'98). Anthony G. Cohn, Lenhard K. Schubert, Stuart C. Shapiro (Eds.). Trento, Italy, june 2-5, 1998. Morgan Kaufmann, 1998.

CHANDRASEKARAN, B.; JOHNSON, T. R.; BENJAMINS, V. R. Ontologies: what are they? why do we need them?. *IEEE Intelligent Systems*, v. 14, n. 1, p. 20-26, 1999.

CHAWATHE, S. et al. TSIMMIS Project: integration of heterogeneous Information Sources. In: PROCEEDINGS OF THE 100th ANNIVERSARY MEETING OF THE INFORMATION PROCESSING SOCIETY OF JAPAN. Tokyo, Japan, p. 7-18, October 1994.

DOMINGUE, J. et al. Supporting Ontology Driven Document Enrichment within Communities of Practice. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON KNOWLEDGE CAPTURE. *International Conference On Knowledge Capture.* Victoria, British Columbia, Canada, 2001.

FENSEL, D. et al. Ontobroker – the very high idea. In: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SOCIETY CONFERENCE. Sanibel Island, Florida, USA, may 18-20, 1998.

GANDON, F. Engineering an Ontology for a Multi-Agents Corporate Memory System. In: PROC. INTERNATIONAL SYMPOSIUM ON THE MANAGEMENT OF INDUSTRIAL AND CORPORATE KNOWLEDGE, p. 209-228, 2001. Disponível em: <http://citeseer.nj.nec.com/gandon01engineering.html>. Acesso em: 22 maio 2002.

GOASDOUÉ, F.; LATTES, V.; ROUSSET, M. C. The Use of CARIN Language and Algorithms for Information Integration: The PICSEL Project. International. *Journal of Cooperative Information Systems* (IJCIS), v. 9, n. 4, p. 383-401, december 2000.

GRUBER, T. (1996). *What is an Ontology?* Disponível em: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>. Acesso em: 14 set 2002.

GRUBER, T. *A translation approach to portable ontology specifications*. London: Academic, 1993. p. 199-220.

GUARINO, N. Understanding, building and using ontologies. In: PROCEEDINGS OF TENTH KNOWLEDGE ACQUISITION FOR KNOWLEDGE-BASED SYSTEMS WORKSHOP, 1996. Disponível em: <http://ksi.cpsc.ucalgary.ca/KAW/KAW96/ guarino/guarino.html#Heading4>. Acesso em: 22 set. 2001.

GUARINO, N.; GIARETTA, P. *Ontologies and KBs, towards a terminological clarification*. In: MARS, N. (Ed.). *Towards a Very Large Knowledge Bases*: Knowledge Building and Knowledge Sharing. [S.l.]: IOS Press, 1995. p. 25-32.

HAAV, H. M.; LUBI, T. L. A survey of concept-based information retrieval tools on the web. In: PROC. OF 5th EAST-EUROPEAN CONFERENCE ADBIS*2001. A. Caplinkas and J. Eder (Eds). *Advances in Databases and Information Systems*, v. 2. p. 29-41, Vilnius "Technika" 2001.

HASMAN, A. et al. (1999). *ID2.1*: Analysis of guideline ontologies. Disponível em: <http://new.euromise.org/mgt/repdev/id21.html>. Acesso em: 29 ago. 2002.

HEFLIN, J.; HENDLER, J. Searching the Web with SHOE. In: ARTIFICIAL INTELLIGENCE FOR WEB SEARCH. *Papers from the AAAI Workshop*. WS-00-01. Menlo Park, AAAI Press, CA, p. 35-40, 2000.

HWANG, C. H. Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information. In: PROCEEDINGS OF THE 6th INTERNATIONAL WORKSHOP ON KNOWLEDGE REPRESENTATION MEETS DATABASES (KRDB'99). Sweden, n. 21, p. 14-20, july 29-30 1999.

JARKE, M.; VASSILIOU, Y. Data Warehouse Quality: A Review of the DWQ Project. In: PROCEEDINGS OF THE 2nd INTERNATIONAL CONFERENCE ON INFORMATION QUALITY (IQ-97). Cambridge, Mass, 1997.

JASPER, R.; USCHOLD, M. A framework for understanding and classifying ontology applications. In: IJCAI-99, ONTOLOGY WORKSHOP. Stockholm, Sweden july 1999.

KALFOGLOU, Y. (2001). *Deploying Ontologies in Software Design*. PhD Thesis. Disponível em: <http://www.ecs.soton.ac.uk/~yk1/research.html>. Acesso em: 21 ago. 2002.

LEGER, A. et al. (2000). *Ontology domain modeling support for multilingual services in e-Commerce*: MKBEEM. Presentation seminar ECAI2000 – Berlin. Disponível em: <http://mkbeem.elibel.tm.fr/paper/ecai00-final.pdf>. Acesso em: 22 dez. 2001.

LUKE, S. et al. Ontology-based Web agents. In: PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS. Marina del Rey, California, United States Publisher ACM Press New York, NY, USA. p. 59-66, 1997.

MARTIN, P. H.; EKLUND, P. (2001). Large-scale cooperatively-built heterogeneous KBs. In: ICCS'01, 9th INTERNATIONAL CONFERENCE ON CONCEPTUAL STRUCTURES. Disponível em:<http://meganesia.int.gu.edu.au/~phmartin/WebKB/ doc/papers/iccs01/>. Acesso em: 6 set. 2002.

MENA, E. et al. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. In: FIRST IFCIS INTERNATIONAL CONFERENCE ON COOPERATIVE INFORMATION SYSTEMS (COOPIS'96). Brussels, Belgium, p. 14-25, june 19-21 1996.

MIZOGUCHI, R.; VANWELKENHUYSEN, J., IKEDA, M. Task ontology for reuse of problem solving knowledge. In: PROC. OF ECAI'94 TOWARDS VERY LARGE KNOWLEDGE BASES. Amsterdam, N. Mars (Ed.), IOS Press, p. 46-59, 1995.

NECHES, R. et al. Enabling technology for knowledge sharing. *AI Magazine*, v. 12, n. 3, Fall 1991.

NODINE, M.; BOHRER, W.; NGU, A. H. Semantic Brokering over Dynamic Heterogeneous Data Sources in InfoSleuth. In: PROCEEDINGS OF THE 15th INTERNATIONAL CONFERENCE ON DATA ENGINEERING. Sydney, Austrialia, IEEE Computer Society, p. 358-365, 23-26 march 1999.

NOY, F. N.; GUINNESS, D. L. (2001). *Ontology development 101*: a guide to create your first ontology. Disponível em: <http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.doc>. Acesso em: 04 maio 2001.

OMELAYENKO, B. Integration of product ontologies for B2B marketplaces: a preview. *ACM: Special Interest Group on Electronic Commerce SIGecom Exchanges*, Newsletter of the ACM SIG on e-commerce, v. 2, n. 1, p. 19-25, 2001.

OUKSEL, A. M.; SHETH, A. Semantic Interoperability. In: SEMANTIC INTEROPERABILITY IN GLOBAL INFORMATION SYSTEMS. Special Section on Semantic Interoperability in Global Information Systems. Simod record Web Edition, v. 28, n. 1, march 1999. Disponível em:< http://www.acm.org/sigmod/record/issues/9903/>.

PREECE, A.; HUI, K. The KRAFT architecture for knowledge fusion and transformation. *Knowledge Based Systems*, v. 13, n. 2-3, p. 113-120, april 2000.

SHETH, A. *Changing focus on interoperability in infomation systems*: from systems, syntax, strutured to semantics, in interoperating geographic operation systems. Norwell, MA M.F.Goodchild, M.J. Egenhofer, R. Fegeas, C. A. Kottman Kluver Pub., 1998.

SHUM, S. B.; MOTTA, E.; DOMINGUE, J. ScholOnto: An Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries*, v. 3. n. 3, p. 237-248, sept. 2000.

STUCKENCHMIDT et al. Methodologies for ontology-based semantic translation. Electronic Commerce Workshop, Brussels, october 2002. Disponível em: <http://www.ecimf.org/events/Brussels-20011016/SemanticTranslation-BUSTER.pdf>.

STUCKENSCHMIDT, H., WACHE, H. Context modeling and transformation for semantic interoperability. In: PROCEEDINGS OF THE 7TH INTERNATIONAL WORKSHOP ON KNOWLEDGE REPRESENTATION MEETS DATABASES (KRDB 2000). Berlin, Germany, august 21, 2000.

USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods an applications. *Knowledge Engineering Review*, v. 11, n. 2, 1996.

VAN HEIJIST, G.; SCHREIBER, A. T.; WIELINGA, B. J. Using Explicit Ontologies in KBS Development. *International Journal of Human-Computer Studies*, v. 46, Issue 2-3, p. 183-192, feb./march 1997.

VÁZQUEZ, E.; VALERA, F.; BELLIDO, L. (2001). *Modelado de Servicios Complejos en uma Plataforma de Intermediación para comercio Electrónico*. Disponível em: <http://www.telecom.ece.ntua.gr/smartec/documentation/Publications/smartec-jitel2001. pdf>. Acesso em: 14 ago. 2002.

WACHE, H. et al. Ontology-based integration of information – a survey of existing approaches. In: IJCAI, WORKSHOP ON ONTOLOGIES AND INFORMATION SHARING, 2001. Disponível em: <http://citeseer.nj.nec.com/wache01ontologybased. html>. Acesso em: 01 maio 2002.

WATSON, A. (2000). *Workshop Object Management Group – OMG*: Creating Interoperability. Disponível em: <http://www.internet2.edu/health/20000710-SURA-AWatson.pdf>. Acesso em: 25 mar. 2002.

WEIHAI, Y. (2002). *Middleware Seminar*. Disponível em: <http://www.cs.uit.no/studier/ kurs/d312/info/2001h/forelesning/middleware.pdf>. Acesso em: 25 mar. 2002.

WIEDERHOLD, G. Mediation to deal with heterogeneous data sources. In: INTERNATIONAL SYMPOSIUM ON FIFTH GENERATION COMPUTER SYSTEMS. Tokyo, Japan, 1994.

* * *