2
Fundamentos

Este capítulo descreve os principais conceitos envolvidos em *alinhamento de catálogos*. A seção 2.1 apresenta a noção de *catálogo* e seus termos correlatos: esquemas, instâncias e tesauros. A seção 2.2 relata algumas das principais abordagens existentes para alinhamento de catálogos. Por fim, a seção 2.3 apresenta três exemplos de aplicações que podem se beneficiar com o uso de técnicas de alinhamento de catálogos.

2.1.
Catálogos de objetos

Existem diversas definições do que seja um catálogo. Segundo o dicionário léxico *WordNet*, um catálogo é "uma lista de coisas, geralmente organizadas sistematicamente" (WordNet, 2007). O glossário da biblioteca da Universidade de Cornell, nos EUA, define um catálogo como "uma compilação de registros descrevendo os conteúdos de uma coleção ou grupo de coleções particulares" (Cornell, 2008). O W3C diz que "catálogos são bancos de dados estruturados com o propósito de visualização e busca e que geralmente são usados para organizar um grande volume de dados relacionados" (W3C, 2008).

Neste trabalho, adotamos a definição de que um catálogo armazena dados sobre um conjunto de objetos de um determinado domínio, tipicamente classificados por algum tipo de taxonomia. Também assumimos que é possível identificar se dois objetos de catálogos heterogêneos distintos representam o mesmo objeto no mundo real.[1] Como exemplos, citamos os catálogos de objetos geográficos (também chamados de *gazetteers*), os catálogos de livros de uma livraria e os catálogos de produtos de uma loja de *e-commerce*.

Para ilustrar essa discussão, podemos considerar como exemplo o *Alexandria Digital Library Gazetteer* (ADL) (ADL, 1999), que é um catálogo de objetos geográficos que possui aproximadamente 5,9 milhões de entradas classificadas de acordo com o ADL Feature Type Thesaurus (FTT), um esquema

[1] Essa informação é importante para os algoritmos de alinhamento, descritos no capítulo 3.

de classificação baseado em um vocabulário controlado. A Tabela 1 exibe um fragmento do catálogo da ADL, onde “class” é o campo utilizado para classificar os objetos.

identifier	name	x	y	class
adlgaz-1-1150831-7c	Valiente, Salto del – Brazil	-53.8	-27.1	Waterfalls
adlgaz-1-1308370-21	Abuna, Rio – Brazil	-65.3833	-9.6833	Streams
adlgaz-1-1457057-00	Rio de Janeiro, Estado do – Brazil	-42.5	-22.0	Administrative áreas
adlgaz-1-1457059-20	Rio de Janeiro, Serra do - Brazil	-44.95	-17.95	Mountains
adlgaz-1-1470793-1b	Vicosa – Brazil	-42.8833	-20.75	Populated places

Tabela 1 – Fragmento de dados do catálogo da ADL

Os três principais elementos que compõem um catálogo são: as *instâncias*, que representam os objetos descritos no catálogo; o *esquema*, que descreve como os dados estão organizados; e um *tesauro* ou *taxonomia*, utilizado para classificar essas instâncias. A seção 2.1.1 define os conceitos de *esquemas* e *instâncias*, enquanto que a seção 2.1.2 faz uma breve discussão sobre taxonomias e tesauros.

2.1.1. Esquemas e instâncias

Em Banco de Dados, um esquema é responsável por prover a descrição de um banco de dados (Elmasri & Navathe, 2006). No esquema de um banco de dados (ou de um catálogo) podem ser encontradas diversas informações, tais como atributos, tipos de dados, relacionamentos, e alguns tipos de restrições. Já as instâncias (ou objetos) representam o conteúdo do banco de dados.

Por simplificação, restringimos o esquema *S* de um catálogo como sendo um conjunto de atributos utilizados para estruturar as informações sobre os objetos de um determinado domínio. Cada atributo representa um aspecto do domínio e possui uma semântica precisa e não-ambígua dentro do domínio em questão. Dizemos que uma instância a_1 *segue* um esquema *S* quando a_1 contém

valores para os atributos de *S*. Como exemplo, a Tabela 2 mostra os principais atributos do esquema do catálogo da ADL, com a semântica de cada um deles.

Atributo	Semântica
identifier	identificador local do objeto
name	nome de exibição do objeto
x	longitude do objeto
y	latitude do objeto
class	termo do FTT utilizado para classificar o objeto

Tabela 2 – Descrição do esquema do catálogo da ADL

Existem diversas maneiras de representar um esquema para uma fonte de dados. A Tabela 3 lista alguns dos tipos mais comuns de representação.

Representação	Descrição
Database schema	O esquema lógico de um banco de dados relacional, gerenciado por um SGBD
XML	Um arquivo XML contendo a descrição do esquema, utilizando o recurso de meta-linguagem
XML Schema	Um arquivo XML contendo a descrição do esquema, utilizando vocabulários controlados de semântica bem definida (do XML Schema)
Grafo	Um conjunto de vértices e arestas com as quais são representados os elementos de um esquema e seus relacionamentos
Ontologia	Uma representação explícita de conceitos, representada em uma linguagem como RDF, DAML, ou OWL

Tabela 3 – Algumas alternativas para representação de esquemas

2.1.2. Taxonomias e tesauros

Originalmente, a palavra “taxonomia” se referia à classificação dos organismos vivos, dada pelo botânico Lineu no século XVIII (Garshol, 2004). Com o passar do tempo, “taxonomia” ganhou a conotação de um sistema para nomeação e organização de objetos em grupos que compartilham características similares (Graef, 2001).

Para nós, uma taxonomia é simplesmente um esquema de classificação que utiliza *termos* (ou *categorias*) para classificar objetos. Taxonomias compõem estruturas essenciais em vários tipos de sistemas de gerenciamento de informação, sejam elas definidas explicita ou implicitamente. Entre as taxonomias mais comuns estão as taxonomias planas e as hierárquicas (Bedford, 2003).

Uma taxonomia plana agrupa os termos num conjunto de categorias sem um relacionamento inerente entre si. Constituem o tipo mais simples de taxonomia. Como exemplo, podemos citar o conjunto de regiões do Brasil (i.e., Norte, Nordeste, Centro-oeste e Sul); o conjunto de classes morfológicas para classificação das palavras de uma sentença (i.e., substantivo, adjetivo, advérbio, verbo, etc.); e o conjunto de menus na barra de ferramentas de uma aplicação[2].

Uma taxonomia hierárquica organiza os termos numa estrutura em árvore. Nesse esquema, o conteúdo é agrupado em dois ou mais níveis, com os relacionamentos entre níveis possuindo significado particular. Relacionamentos entre categoria-pai e categoria-filha podem significar pertinência ou especialização. Como exemplo, podemos citar os diretórios de um sistema de arquivos de um sistema operacional; e a classificação dos seres vivos do reino Animália. Esta última aparece ilustrada, de maneira simplificada, na Figura 1.

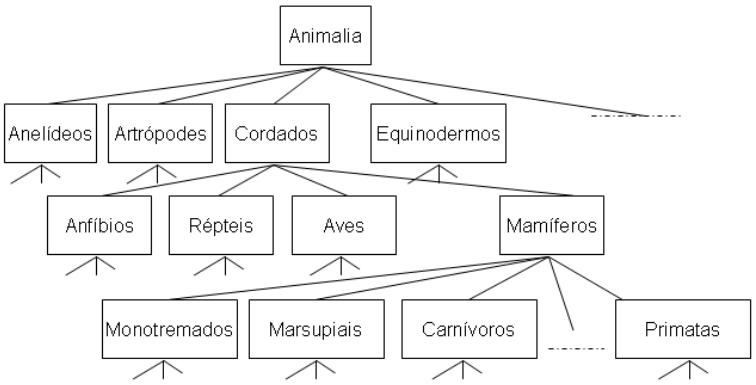

Figura 1 – Taxonomia hierárquica do reino Animália

Segundo a UNESCO, um tesauro é “uma lista estruturada e definida de termos que padroniza palavras utilizadas para indexação“ (Unesco, 1995). De

[2] Nesses exemplos os termos podem possuir alguns relacionamentos entre si. Por exemplo, as regiões do Brasil possuem uma relação de adjacência. No entanto, por não termos interesse nesses relacionamentos, podemos modelar esses casos como taxonomias planas.

maneira equivalente, um tesauros pode ser visto como um vocabulário de uma linguagem de indexação controlada, formalizado de maneira a deixar explícitos os relacionamentos entre os termos (ISO 2788, 1986). O padrão ISO2788 define seis tipos de propriedades para cada termo de um tesauro (cinco das quais expressam relacionamentos entre termos). A Tabela 4 exibe essas propriedades.

Propriedade	Abreviação	Descrição
Scope Note	SN	É uma cadeia de caracteres que explica o significado do termo ao qual está associada
Use	USE	O termo que se segue ao símbolo é o termo preferido quando é utilizado como sinônimo
Use For	UF	Inverso da propriedade USE
Top Term	TT	O termo que se segue ao símbolo é o termo mais geral ao qual o termo anterior ao símbolo pertence
Broader Term	BT	O termo que se segue ao símbolo é um termo de significado mais amplo do que o termo anterior ao símbolo
Narrower Term	NT	O termo que se segue ao símbolo é um termo de significado mais específico do que o termo anterior ao símbolo
Related Term	RT	O termo que se segue ao símbolo é um termo associado ao termo anterior ao símbolo mas não é um sinônimo ou quase-sinônimo

Tabela 4 – Propriedades de termos de um tesauro

Como um exemplo de tesauro, podemos citar o Feature Type Thesaurus (FFT) da ADL. Um fragmento do FFT está exibido na Figura 2.

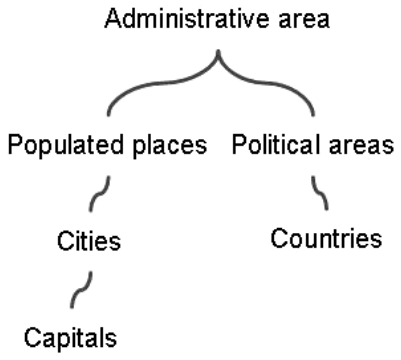

Figura 2 – Fragmento do FTT

Neste trabalho, a menos que explicitamente indicado, utilizaremos genericamente o termo "tesauro" [3] para nos referir ao esquema de classificação de um catálogo, seja ele uma simples taxonomia plana ou um tesauro propriamente dito.

2.2. Alinhamento de catálogos

2.2.1. Definições básicas

Chamamos de *alinhamento* à determinação de uma correspondência ≈ entre os elementos de interesse de um catálogo C_1 com os de um outro catálogo C_2. Chamamos o catálogo C_1 de *catálogo-fonte* e o catálogo C_2 de *catálogo-alvo*. Também dizemos que um elemento e_1 de C_1 *alinha* com um elemento e_2 de C_2 sse $e_1 \approx e_2$.

Neste trabalho, dividimos o alinhamento de catálogos em dois tipos:

- Alinhamento de *esquemas*: é o alinhamento em que os elementos de interesse são os atributos dos esquemas dos catálogos considerados. Ou seja, o objetivo é encontrar os mapeamentos dos atributos de um *esquema-fonte* S_1 (pertencente a um catálogo C_1) para os atributos de um *esquema-alvo* S_2 (pertencente a um catálogo C_2).
- Alinhamento de *tesauros*: é o alinhamento em que os elementos de interesse são os termos dos tesauros dos catálogos considerados. Ou seja, o objetivo é encontrar os mapeamentos dos termos de um *tesauro-fonte* T_1 (pertencente a um catálogo C_1) para os termos de um *tesauro-alvo* T_2 (pertencente a um catálogo C_2).

Em um contexto real, alinhamentos podem ter diferentes cardinalidades. Dados dois esquemas[4] S_1 e S_2, podemos ter alinhamentos 1:1 (associação de um único atributo de S_1 a um único atributo de S_2, ou atributos sem nenhuma associação), 1:n (associação de um único atributo de S_1 a vários atributos de S_2), m:1 (associação de vários atributos de S_1 a um único atributo de S_2) ou m:n

[3] O dicionário Aurélio define taxonomia como: "1. Ciência da Classificação; 2. (Biol.) Sistemática; 3. (E.Ling.) Classificação das palavras". Como esses significados diferem do significado utilizado neste texto, optamos por empregar o termo "tesauro" como sinônimo de "taxonomia", por haver maior consenso quanto ao seu significado.

[4] A explicação é análoga para o alinhamento entre os termos de dois tesauros.

(associação de vários atributos de S_1 a vários atributos de S_2). (Bilke, 2007) classifica alinhamentos do tipo 1:1 como alinhamentos *simples* e alinhamentos dos tipos 1:n, m:1 e m:n como alinhamentos *complexos*.

Como exemplo, podemos considerar a Figura 3, a qual ilustra o alinhamento de dois catálogos fictícios contendo informações de funcionários de uma empresa do setor de Tecnologia da Informação. O exemplo ilustra tanto o alinhamento de esquemas, quanto o alinhamento de tesauros[5].

Neste trabalho, quando dizemos que dois elementos alinham, apenas informamos que os mesmos possuem um determinado *grau de similaridade*. Ademais, nos limitamos ao tratamento de alinhamentos simples de catálogos. Dessa forma, alinhamentos de maior cardinalidade, como {*nome*, *sobrenome*} ≈ *nome* e *endereço* ≈ {*logradouro*, *bairro*}, exibidos na Figura 3, não são considerados.

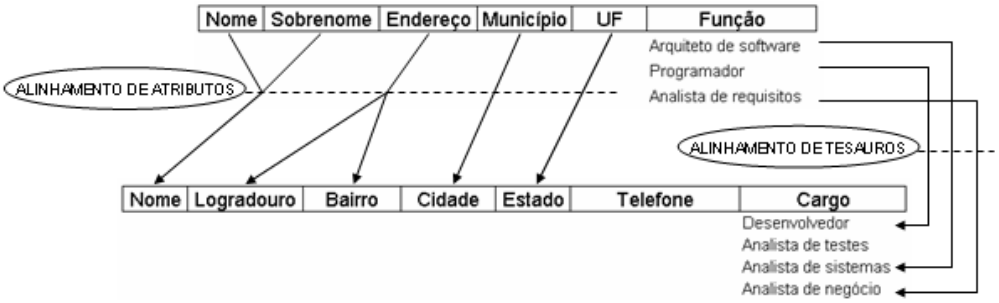

Figura 3 – Ilustração do alinhamento de dois catálogos fictícios

O uso de catálogos como alvos para técnicas de alinhamento, em contraste com o uso de bancos de dados de estrutura arbitrária, permite que possamos nos abstrair de problemas relacionados à heterogeneidade estrutural, focando unicamente no problema de heterogeneidade semântica. Essa simplificação é prática e particularmente útil em aplicações de mediadores, cujas fontes de dados podem ser vistas como catálogos com esquemas equivalentes aos respectivos esquemas exportados (seção 2.3.3).

Existem diferentes abordagens que podem ser utilizadas para realizar o alinhamento de catálogos. Rahm & Bernstein (2001) apresentam uma classificação bastante completa de algumas abordagens para alinhamento de esquemas, fazendo uso de diversos critérios. Na escolha de uma técnica para

[5] Neste caso, temos uma simples taxonomia de classificação de funcionários quanto ao cargo ocupado na empresa.

resolver um determinado problema, pode-se utilizar uma abordagem seguindo um único critério ou combinarem-se abordagens que usam diversos critérios. Esta última costuma produzir resultados melhores. A classificação de Rahm & Bernstein (2001) pode ser visualizada na Figura 4.

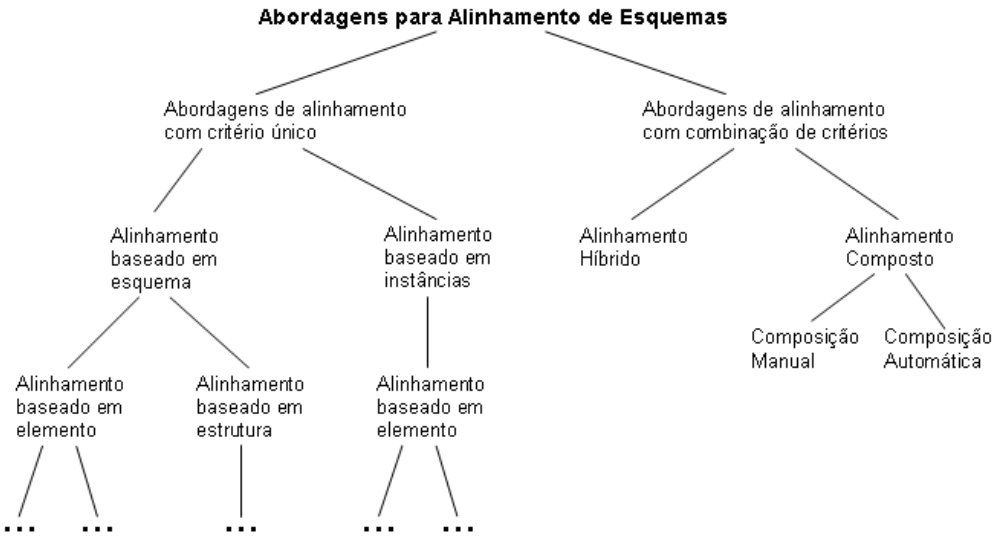

Figura 4 – Uma classificação de abordagens para alinhamento de esquemas (Rahm & Bernstein, 2001)

Como pode ser visto, as abordagens presentes na classificação exibida na Figura 4 podem ser organizadas em dois grandes grupos: abordagens *baseada em esquema* (sintáticas) e abordagens *baseadas em instâncias* (semântica). Ambas consideram o problema de alinhamento de esquemas pré-existentes, ou seja, são abordagens *a posteriori*. Além dessas abordagens, Casanova et al. (2007) destacam técnicas que podem ser utilizadas para a realização de alinhamentos *a priori*, ou seja, quando da concepção dos esquemas e catálogos.

A Figura 5 resume os grandes grupos de técnicas de alinhamento identificados por Casanova et al. (2007), os quais são descritos em mais detalhes nas seções que se seguem.

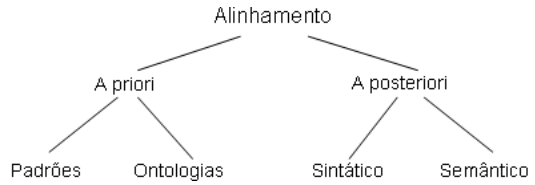

Figura 5 – Grandes grupos de abordagens para alinhamento (Casanova et al., 2007)

A seção 2.2.2 explica os fundamentos do alinhamento baseado em esquema (sintático). A seção 2.2.3 descreve o alinhamento baseado em instâncias (semântico); enquanto que a seção 2.2.4. explica o alinhamento a priori, baseado em padrões e ontologias.

2.2.2. Alinhamento baseado em esquema

O alinhamento *sintático*, ou *baseado em esquema*, considera apenas informações em nível de esquema ou metadados, ou seja, não são utilizadas informações relativas a instâncias. Entre as informações disponíveis que podem ser utilizadas estão os nomes, tipos, restrições e estruturas dos elementos que compõem o esquema (ou taxonomia).

Uma abordagem comum para alinhamento baseado em esquema consiste em utilizar uma medida de similaridade sintática, assumindo que quanto mais próximo sintaticamente estiverem dois elementos, mais próximos semanticamente esses elementos estarão.

Como um exemplo, considere os esquemas exibidos na Figura 6, pertencentes a dois catálogos do domínio de filmes, alinhados utilizando uma abordagem sintática. Nesse caso, os atributos *Título* e *Direção* de ambos os esquemas estão corretamente alinhados. No entanto, os atributos *Classificação* dos dois esquemas possuem diferentes significados em cada um dos catálogos. No primeiro esquema, *Classificação* se refere à faixa etária recomendada de acordo com o conteúdo do filme, ao passo que o atributo *Classificação* do segundo esquema se refere ao resultado da avaliação dos filmes por um grupo de especialistas. Conseqüentemente, é inválido o alinhamento encontrado.

Em resumo, a abordagem sintática não é suficientemente prática, por não ser efetiva nos seguintes casos:

- **Ocorrência de heterogeneidade sintática entre as fontes**: quando designações diferentes são utilizadas para nomear os mesmos conceitos, a similaridade sintática entre os termos pouco informará. Diferenças como idioma, uso de siglas, nomes de atributos pouco significativos, etc. tornam a abordagem pouco efetiva.
- **Ocorrência de heterogeneidade semântica entre as fontes:** nem sempre designações idênticas são utilizadas para nomear os mesmos conceitos. Em outras palavras, proximidade sintática não implica

proximidade semântica (Casanova et al., 2007) (caso dos dois atributos chamados *Classificação*, no exemplo da Figura 6).

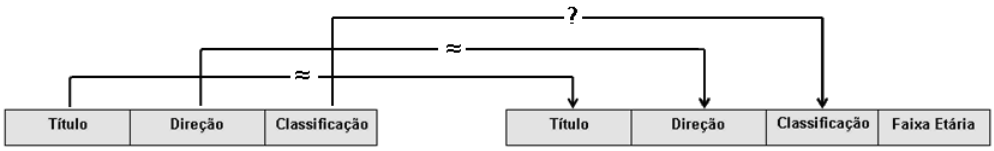

Figura 6 – Alinhamento sintático de esquemas

2.2.3. Alinhamento baseado em instâncias

O alinhamento *semântico*, ou *baseado em instâncias*, considera informações obtidas a partir das instâncias de um catálogo para inferir os alinhamentos. Baseia-se na premissa de que os dados podem prover informações importantes sobre o conteúdo e o significado dos elementos dos esquemas (Rahm & Bernstein, 2001).

Retornando ao exemplo da seção anterior, podemos utilizar um conjunto de instâncias típicas encontradas num dos catálogos, como “Coração valente” e “Rocky Balboa”, para verificar a ocorrência dessas instâncias no outro catálogo. Ao compararmos essas instâncias, além do alinhamento dos atributos *Título* e *Direção*, identificamos corretamente o alinhamento entre *Classificação*, no primeiro esquema, com *Faixa Etária*, no segundo esquema, conforme ilustra a Figura 7.

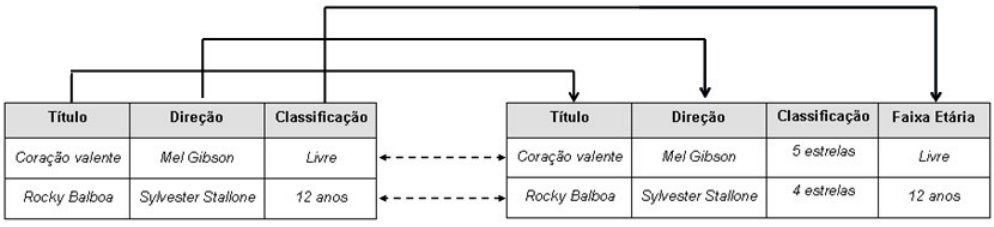

Figura 7 – Alinhamento semântico de esquemas

Além de os catálogos estarem inseridos num mesmo domínio, para empregar a abordagem semântica na prática, é necessário que se possa determinar se dois objetos de diferentes catálogos são iguais, isto é, representam o mesmo objeto no mundo real (Casanova et al., 2007). Em vários domínios, existem identificadores únicos que podem ser usados com esse

propósito, como o ISBN dos livros, os endereços IP dos computadores em uma determinada rede, o número de CPF dos cidadãos brasileiros, etc.

Entretanto, a abordagem baseada em instâncias pode apresentar problemas mesmo no alinhamento de catálogos situados num mesmo domínio, devido à heterogeneidade semântica *no nível de instâncias*. Considerando o primeiro catálogo do exemplo anterior, sabemos que, além de ter sido diretor do filme "Coração valente", Mel Gibson também interpretou o protagonista do filme. O mesmo pode ser afirmado sobre Sylvester Stallone com relação ao filme Rocky Balboa. Sendo assim, se, no primeiro esquema, em vez de representarmos a informação a respeito da *Direção*, quiséssemos representar a informação de *Ator protagonista* do filme, teríamos o alinhamento de *Ator protagonista*, no primeiro esquema, com *Direção*, no segundo esquema, o que, obviamente, não é o caso.

Outro problema com o emprego de instâncias ocorre quando trabalhamos com domínios cujos dados se alteram frequentemente ao longo do tempo. Nessa situação, instâncias podem representar o mesmo conceito do ponto de vista semântico, mas não terem os mesmo valores para alguns de seus atributos, devido à heterogeneidade temporal. Como exemplos, podemos considerar o alinhamento de diversos catálogos contendo dados estatísticos de regiões geográficas ao longo dos últimos vinte anos, ou catálogos contendo sistemas de software em versões variadas.

2.2.4. Alinhamento baseado em padrões e ontologias

Em contraste com as abordagens sintática e semântica discutidas acima, que são abordagens a posteriori, Casanova et al. (2007) propõem uma abordagem a priori para endereçar o problema do alinhamento. A idéia é que, ao receber a tarefa de especificar um novo catálogo (ou banco de dados), o projetista primeiramente selecione um padrão apropriado que possa guiá-lo na confecção dos esquemas e tesauros de seu catálogo. Caso não exista um padrão apropriado, o projetista poderia submeter uma proposta para um esquema ou taxonomia cobrindo o domínio de aplicação em questão. Alinhar os elementos de dois esquemas projetados seguindo um mesmo padrão torna-se tarefa trivial, visto que existe um consenso quanto à semântica de cada elemento do esquema, evitando-se possíveis ambigüidades.

Existem diversas organizações empenhadas na elaboração de padrões para os mais variados domínios de aplicação, como a *Energistics* (padrões para a área de Energia e Petróleo), o *Open Geospatial Consortium* (padrões para dados geográficos), e o IEEE (padrões para telecomunicações, informática e geração de energia). Casanova et al. (2007) destacam o padrão ISO 19115 (2003), que define esquemas de metadados para descrição de objetos geográficos. Esses esquemas de metadados possuem um núcleo de elementos comuns e um núcleo de elementos opcionais. Cada aplicação deve instanciar o núcleo comum e especificar quais são os elementos opcionais que implementa, definindo, assim, um perfil (ou "*profile*"). Com isso, podem-se alinhar catálogos de objetos geográficos que obedeçam ao padrão, bastando-se ter o conhecimento do perfil implementado por cada um dos catálogos.

Nesta abordagem, se nenhum padrão existir, o projetista do catálogo pode tentar selecionar fragmentos de ontologias de nível-superior que contenham os conceitos pertencentes ao domínio de aplicação (Casanova et al., 2007). Os autores enumeram as seguintes etapas para a definição de um esquema comum para um domínio de aplicação:

- Selecionar fragmentos de ontologias conhecidas que cubram conceitos pertencentes ao domínio em questão;
- Alinhar elementos de diferentes fragmentos em conceitos unificados;
- Publicar os conceitos unificados como uma ontologia, indicando quais conceitos são obrigatórios e quais são opcionais.

Existem diversas ontologias que podem ser utilizadas como base para essa estratégia. Como exemplo, podemos citar OpenCyc(OPENCYC), a *Suggested Upper Merged Ontology* (SUMO) (SUMO, 2006; Niles & Pease, 2001), e a Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Masolo et al., 2003).

Devido à falta de semântica formal na especificação dos bancos de dados existentes, torna-se bastante difícil automatizar completamente as técnicas de alinhamento a posteriori. Daí a importância das idéias propostas em uma abordagem a priori, como discutido por Casanova et al. (2007). No entanto, como descrito em Moulton (2002), o uso prático de padrões ou ontologias encontram diversas dificuldades. Primeiro, a existência de padrões concorrentes em um mesmo domínio ou mesmo múltiplas versões de um mesmo padrão podem inviabilizar a estratégia. Segundo, existe a obrigatoriedade de adoção do

padrão por todos os projetistas, o que nem sempre é o caso. Por último, mesmo com a adoção de um único padrão, pode haver diferentes interpretações do mesmo, ocasionando problemas semânticos.

Optamos por explorar abordagens baseadas em instâncias para implementação em nossa infra-estrutura de software, devido à ineficácia da abordagem sintática e aos entraves à estratégia a priori.

2.3. Aplicações de alinhamento de catálogos

Nesta seção, apresentamos três exemplos de aplicações para as quais se torna importante o emprego de técnicas de alinhamento. Na seção 2.3.1 discorremos sobre *bancos de dados federados*, na seção 2.3.2 abordamos sistemas de *data warehouse* e, por fim, destacamos aplicações de *mediadores*, na seção 2.3.3.

2.3.1. Sistemas de bancos de dados federados

Um sistema de banco de dados federados (SBDF) é composto por um conjunto de sistemas de bancos de dados autônomos, e possivelmente heterogêneos, que cooperam entre si (Sheth & Larson, 1990). Para administrar o sistema, pode-se ter um componente que mantenha a topologia da federação e que supervisione a entrada de novos bancos de dados (sistema gerenciador de bancos de dados federados, ou SGBDF). Usualmente, cada sistema de banco de dados deve definir o seu *esquema de importação* e o seu *esquema de exportação* (Sheth & Larson, 1990). O esquema de importação especifica quais são as informações que o sistema precisa obter dos demais sistemas na federação. O esquema de exportação define as informações e serviços que um sistema de banco de dados disponibiliza para os demais sistemas da federação.

Numa federação típica, é necessário que cada sistema tenha a capacidade de se comunicar com todos os demais. Como ilustra a Figura 8, o problema com essa arquitetura é que, se existem *N* sistemas de bancos de dados, então haverá *N – 1* comunicações entre cada um deles, o que tornará necessária a escrita de *N – 1* trechos de código para suportar as consultas entre sistemas (Garcia-Molina et al., 2000). Para amenizar o problema, poderíamos embutir um módulo de alinhamento de esquemas no SGBDF, o qual se encarregaria de determinar o alinhamento dos esquemas exportados de cada um dos sistemas

de bancos de dados registrados na federação. Com isso, quando um sistema *A* da federação precisasse se comunicar com um sistema *B*, bastaria o sistema de *A* se comunicar com o SGBDF para obter as informações de alinhamento entre os esquemas dos bancos de dados *A* e *B*.

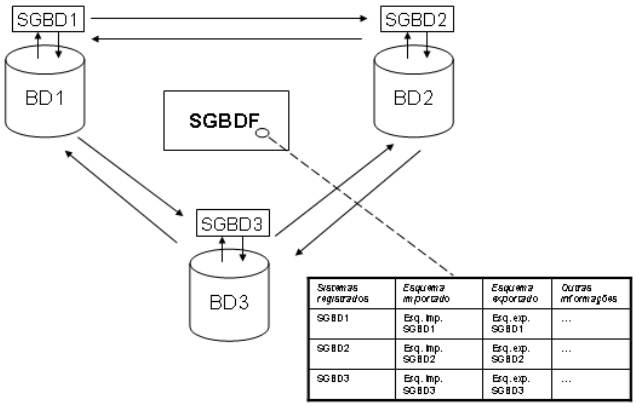

Figura 8 – Um típico sistema de banco de dados federados

2.3.2. Data warehouses

Uma Data Warehouse (DW) é um grande banco de dados com dados oriundos de diversos bancos de dados. As DWs geralmente são utilizadas em conjunto com ferramentas de processamento analítico de consultas (conhecidas como ferramentas *OLAP*). O objetivo primário desses sistemas é fornecer suporte à tomada de decisão através da geração de relatórios e gráficos obtidos a partir do processamento dos dados. O uso de algoritmos de mineração de dados e um modelo de dados multidimensional também caracterizam a maioria desses sistemas.

Um sistema de DW é visto pelo usuário como um único sistema. Consultas podem ser emitidas pelos usuários exatamente como seriam feitas a um sistema de banco de dados comum. No entanto, as consultas são geralmente restritas a operações de leitura dos dados para que a DW possa se manter num estado consistente com os BDs de origem.

Depois de ser populada, uma DW deve ter suas informações atualizadas de tempos em tempos para refletir atualizações feitas nos bancos de dados de origem. Existem pelo menos três abordagens para a manutenção de uma DW (Garcia-Molina et al., 2000):

- A DW é periodicamente reconstruída a partir dos dados correntes nos bancos de dados de origem;

- A DW é atualizada periodicamente, de acordo com as mudanças realizadas nos bancos de dados de origem, desde o último carregamento de dados;
- A DW é atualizada imediatamente sempre que alguma mudança ocorre em algum dos bancos de dados de origem.

Uma arquitetura típica de uma data warehouse por ser visualizada na Figura 9 (Özsu & Valduriez, 1999). Várias operações devem ser realizadas durante o ciclo de vida de uma data warehouse, como *extração de dados*, *transformação* e *carregamento* (Chaudhuri & Dayal, 1997). A DW deve possuir um esquema global que captura as informações comuns pertencentes aos esquemas dos bancos de dados de origem. Com isso, é necessário mapear os esquemas dos BDs de origem no esquema da DW. Novamente, um módulo de integração que faça uso de técnicas de alinhamento pode facilitar bastante essa tarefa.

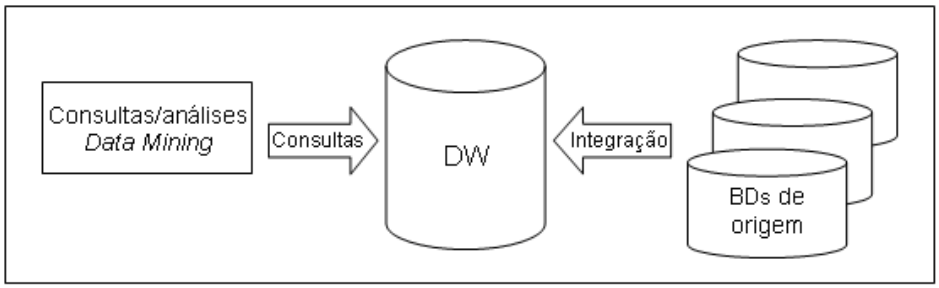

Figura 9 – Arquitetura de uma DW

2.3.3. Mediadores

Um mediador é um componente de software que facilita o acesso a um conjunto de fontes de dados (Wiederhold, 1992). É tarefa do mediador receber consultas dos usuários, traduzir essas consultas em consultas que podem ser entendidas por cada fonte, receber e combinar os resultados provenientes de cada fonte, devolvendo-os ao usuário.

Cada fonte de dados registrada no mediador deve definir seu esquema de exportação, isto é, o conjunto de atributos que deseja tornar disponível para consulta aos usuários. Além disso, o próprio mediador deve definir um esquema global (ou esquema mediado) o qual é semelhante a uma visão (virtual) que integra as diferentes fontes de dados registradas. O esquema global de um mediador é semelhante ao esquema global da DW. A diferenca está em que,

nesta última, os dados estão materializados (Garcia-Molina et al., 2000). Diferentemente dos sistemas de bancos de dados federados e das data warehouses, os quais assumem uma interface SQL para o sistema, uma aplicação de mediação de consultas lida com diferentes fontes de dados, as quais podem ir desde simples arquivos até SGBDs completos (Garcia-Molina et al., 2000).

Para endereçar esses problemas, a comunidade de banco de dados propôs uma arquitetura de adaptadores (*wrappers*) para fontes de dados (Özsu & Valduriez, 1999). A função de um adaptador é exportar dados e serviços da fonte de dados da qual está encarregado. A Figura 10 mostra um sistema de mediação de consultas implementado utilizando a arquitetura mediador-adaptador. Nesse exemplo, há a integração de três tipos de fontes de dados: um banco de dados comum, um arquivo XML e um *Web Service*. Num sistema desse tipo, os usuários emitem consultas ao mediador e este, por sua vez, consulta cada uma das fontes por meio dos adaptadores, os quais possuem a função de traduzir uma consulta de alto nível especificada pelo mediador para a tecnologia específica da fonte de dados em questão. Após reunir os resultados fornecidos por todas as fontes, o mediador combina esses resultados e os retorna para o usuário.

Não existe um consenso em quais dados e serviços os adaptadores devem expor a respeito de uma fonte e nem como isso deve ser feito (Özsu & Valduriez, 1999). Mesmo assim, o modelo mediador-adaptador é uma abstração amplamente adotada para o problema da integracao de fontes de dados (um mediador pode ser usado como um banco de dados de origem para uma DW).

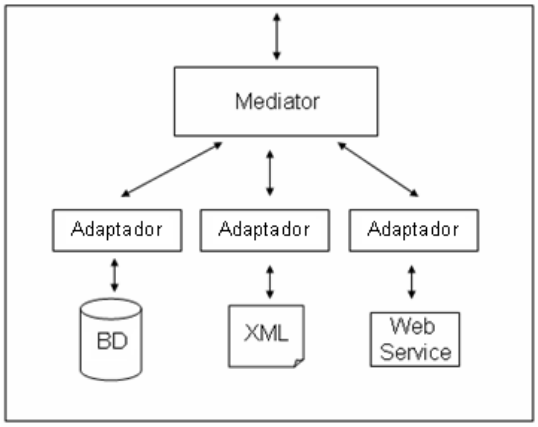

Figura 10 – Sistema implementando a arquitetura mediador-adaptador

A arquitetura mediator-adaptador possui diversas vantagens (Özsu & Valduriez, 1999). Componentes especializados da arquitetura permitem que os diversos interesses dos diferentes tipos de usuários possam ser tratados separadamente; além disso, mediadores tipicamente se especializam num conjunto de fontes de dados relacionadas, as quais possuem dados similares, e, por isso, exportam esquemas e semântica considerando um dominio particular. A especialização dos componentes leva a um sistema distribuído flexível e extensível.

Aplicações de mediadores são bastante promissoras, devido à explosão de crescimento da World Wide Web (Web). O sucesso da Web deve-se, entre outras coisas, à facilidade de publicação de dados e de acesso a esses dados provida pelas máquinas de busca. No entanto, as máquinas de busca atuais ainda são bastante limitadas, por apresentar capacidades de cobertura e consulta limitadas (Casanova, 2007). Uma proposta para explorar de maneira mais completa toda a informação que a Web pode oferecer é construir Cooperativas de Fontes de Dados, as quais possuem mediadores como componentes centrais.

Para que os mediadores possam converter as consultas dos usuários de maneira apropriada a cada fonte de dados, é fundamental que o mediador tenha um mapeamento de seu esquema mediado nos esquemas de cada uma das fontes registradas. Novamente, é de grande importância automatizar ao máximo esse processo, provendo ao sistema de mediação um módulo capaz de realizar o alinhamento do esquema mediado com os esquemas exportados por cada fonte de dados.

Apesar de serem aplicáveis em diversos tipos de sistemas, as técnicas de alinhamento de esquemas e tesauros implementadas na infra-estrutura de software descrita no capítulo seguinte foram escolhidas e desenvolvidas visando seu uso por mediadores, dada à relevância destes no contexto atual da Web. O capítulo 4 apresenta um mediador na Web para catálogos.