9 Integração e interoperabilidade entre fontes de dados geográficos

Marco Antonio Casanova
Daniela Francisco Brauner
Gilberto Câmara
Paulo de Oliveira Lima Júnior

9.1 Introdução

Este capítulo aborda inicialmente o problema básico de intercâmbio de dados geográficos. Em seguida, apresenta um breve resumo da arquitetura e das tecnologias relacionadas a integração e interoperabilidade no contexto de dados geográficos, ilustrado com exemplos. Focaliza então a definição de mapeamentos entre fontes de dados. Um exemplo simples ilustra as dificuldades encontradas para criar mapeamentos entre fontes que apresentem heterogeneidade nos dados, tanto em formato e estrutura, quanto em interpretação ou significado. Por fim, o capítulo trata da construção de mediadores capazes de distribuir consultas entre diversas fontes e combinar os resultados devolvidos em uma única resposta para o usuário.

O Capítulo 10 tratará em detalhe as questões de interoperabilidade especificas para o contexto da Internet, enquanto que o Capítulo 11 resumirá as propostas do *Open Geospatial Consortium* (OGC) endereçando interoperabilidade.

A motivação para este e os próximos dois capítulos nasce do crescente volume de fontes independentes de dados geográficos, interligadas entre si, fato que abre uma oportunidade única para intercâmbio de dados em tempo hábil, reduzindo custos e agilizando processos de decisão. No

entanto, para tal, as aplicações devem ser capazes de interpretar e processar dados oriundos de diversas fontes, bem como localizar e acessar as próprias fontes.

Referências para problemas específicos sobre integração e interoperabilidade entre fontes de dados geográficos encontram-se ao longo do texto e sugestões para leitura adicional, ao final do capítulo.

9.2 Intercâmbio de dados geográficos

9.2.1 Problemas inerentes ao intercâmbio de dados geográficos

Entendemos por intercâmbio de dados a capacidade de compartilhar e trocar informações e processos entre diferentes usuários de informação. O grande desafio do intercâmbio de dados é enfrentar a diversidade de modelos conceituais[1] dos SIGs disponíveis no mercado. Esta diversidade faz com que muitas organizações produtoras de informação georreferenciada sigam regras conceituais vinculadas ao sistema por elas utilizado. O resultado é um ambiente heterogêneo, onde cada organização tem sua maneira de organizar a informação espacial.

A falta de modelos conceituais comuns acarreta problemas na troca de dados entre organizações utilizando SIGs distintos. Estes problemas incluem distorção de dados, perdas de qualidade da informação e de definições de atributos e informação sobre georreferenciamento. A tarefa de compartilhamento de dados geográficos deve envolver processos para garantir que a informação não seja perdida ou corrompida na transferência, e ferramentas para prevenir inconsistências resultantes de conjuntos de dados redundantes. Dada a variedade de usuários e diversidade do uso do dado espacial e sistemas de computação, é conveniente dispor de mecanismos de intercâmbio para compartilhar dados entre diferentes sistemas de computação.

Em SIGs, realizar intercâmbio de dados não é uma tarefa simples, devido a complexidade da informação geográfica envolvida, ocorrendo incompatibilidades em vários níveis. O problema vem sendo estudado em

---

[1] Segundo Thomé (1998) entende-se por modelos conceituais a semântica do funcionamento de cada SIG, e a maneira como os dados devem estar organizados.

diferentes níveis, como a conversão entre formatos de dados próprios de cada SIG, a conversão entre semânticas de bancos de dados distintos e o desenvolvimento ou uso de modelos gerais de dados geográficos propostos por diferentes organizações.

9.2.2 Conversão sintática de dados geográficos

O armazenamento dos dados geográficos em um SIG é organizado em estruturas próprias que descrevem características dos dados, por exemplo, coordenadas dos pontos que formam um polígono representando geometricamente uma dada entidade geográfica. As entidades geográficas possuem uma representação geométrica ou geometria e atributos associados. A geometria vetorial é baseada nas primitivas: ponto, linha e polígonos, que podem ser derivadas para formar estruturas mais complexas. A Figura 9.1 mostra exemplos de geometrias vetoriais.

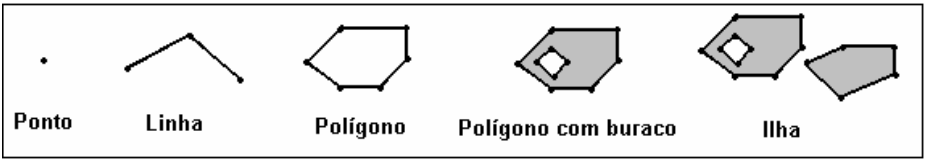
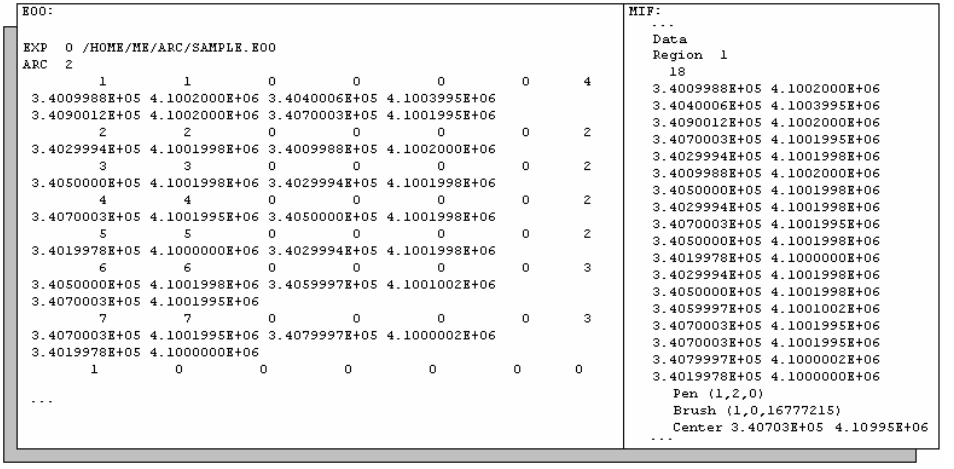


Figura 9.1 – Geometrias vetoriais usadas em SIGs.

Normalmente a organização dos dados nos SIGs é distribuída em "camadas" ("*layers*" ou "planos de informação"), onde cada camada contém uma variável geográfica distinta, por exemplo uma imagem de satélite de uma região, os municípios desta região, a sua geomorfologia, ou hidrologia. Cada camada é representada internamente em estruturas lógicas próprias de cada SIG e armazenada em arquivos distintos de acordo com o formato próprio do sistema utilizado. Este esquema de codificação e os arquivos de sistema ou de exportação possuem uma sintaxe própria para descrever as entidades (geometria e atributos), ou seja, a forma de escrever o dado. Consideramos este esquema próprio de cada sistema para armazenar e documentar seus dados, o nível sintático.

Tradicionalmente, muito do trabalho realizado em intercâmbio de dados geográficos trata de transformações estruturais (Gardels, 1996). A abordagem mais básica é a conversão sintática direta de formatos, que

procura realizar a interpretação e tradução dos arquivos de informação geográfica em diferentes formatos, permitindo que um sistema compreenda os dados provenientes de outros sistemas. Isto é eficiente desde que o desenvolvedor da conversão tenha conhecimento dos formatos envolvidos para não comprometer a qualidade dos dados no processo de conversão. Cada sistema tem sua própria definição e nomenclatura para as diferentes formas de geometria. A Figura 9.2 mostra dois trechos de arquivos em diferentes formatos de exportação. O lado esquerdo é um fragmento de um arquivo com extensão .E00 proveniente do software Arc/Info o da direita é um fragmento de um arquivo de extensão .MIF proveniente do software MapInfo. Ambos descrevem uma mesma entidade, com a mesma geometria e as mesmas coordenadas, mas com uma sintaxe própria.



```
E00:

EXP  0 /HOME/ME/ARC/SAMPLE.E00
ARC  2
        1         1         0         0         0         0       4
 3.4009988E+05 4.1002000E+06 3.4040006E+05 4.1003995E+06
 3.4090012E+05 4.1002000E+06 3.4070003E+05 4.1001995E+06
        2         2         0         0         0         0       2
 3.4029994E+05 4.1001998E+06 3.4009988E+05 4.1002000E+06
        3         3         0         0         0         0       2
 3.4050000E+05 4.1001998E+06 3.4029994E+05 4.1001998E+06
        4         4         0         0         0         0       2
 3.4070003E+05 4.1001995E+06 3.4050000E+05 4.1001998E+06
        5         5         0         0         0         0       2
 3.4019978E+05 4.1000000E+06 3.4029994E+05 4.1001998E+06
        6         6         0         0         0         0       3
 3.4050000E+05 4.1001998E+06 3.4059997E+05 4.1001002E+06
 3.4070003E+05 4.1001995E+06
        7         7         0         0         0         0       3
 3.4070003E+05 4.1001995E+06 3.4079997E+05 4.1000002E+06
 3.4019978E+05 4.1000000E+06
       1         0         0         0         0         0       0
...
```

```
MIF:
  ...
  Data
  Region  1
    18
  3.4009988E+05 4.1002000E+06
  3.4040006E+05 4.1003995E+06
  3.4090012E+05 4.1002000E+06
  3.4070003E+05 4.1001995E+06
  3.4029994E+05 4.1001998E+06
  3.4009988E+05 4.1002000E+06
  3.4050000E+05 4.1001998E+06
  3.4029994E+05 4.1001998E+06
  3.4070003E+05 4.1001995E+06
  3.4050000E+05 4.1001998E+06
  3.4019978E+05 4.1000000E+06
  3.4029994E+05 4.1001998E+06
  3.4050000E+05 4.1001998E+06
  3.4059997E+05 4.1001002E+06
  3.4070003E+05 4.1001995E+06
  3.4070003E+05 4.1001995E+06
  3.4079997E+05 4.1000002E+06
  3.4019978E+05 4.1000000E+06
    Pen (1,2,0)
    Brush (1,0,16777215)
    Center 3.40703E+05 4.10995E+06
  ...
```

Figura 9.2 – Arquivos diferentes - E00 x MIF.

Entendendo a estrutura específica de um formato, é possível escrever um código que trata as características de cada sistema envolvido, viabilizando a conversão ou importação direta, atingindo desta forma um grau de interoperabilidade no nível sintático.

Por fim, cabe salientar que, apesar dos avanços no uso de gerenciadores de dados geográficos, a primeira geração de SIGs possui suporte limitado a

banco de dados e utiliza arquivos para armazenamento e exportação de dados espaciais, o que torna possível encontrar um acervo relevante de dados espaciais em arquivos de diferentes formatos. Assim, a abordagem mais básica para intercâmbio de dados geográficos é a conversão sintática direta, que procura realizar a tradução dos arquivos de informação geográfica entre diferentes formatos.

Para permitir este tipo de conversão, os SIGs trabalham com duas alternativas:

- Oferecer um formato de exportação ASCII de fácil legibilidade, como DXF (Autocad), MID/MIF (MapInfo), E00 (Arc/Info) e SPR (Spring);
- Documentar as estruturas de dados internas, como é o caso do SHP (ArcView).

9.2.3 Uso de metadados

Metadados são "dados sobre os dados", descrevem o conteúdo, condição, histórico, localização e outras características do dado, (FGDC, 2001). O objetivo do seu uso é ter um mecanismo para identificar qual dado existe, a sua qualidade, como acessá-lo e usá-lo. Assim, os metadados tratam a interoperabilidade em nível de gerenciamento da informação, facilitando a recuperação de uma informação contida em um banco de dados.

Por exemplo, uma base de dados contendo mapas com a informação sobre aptidão climática ao plantio de várias culturas pode incluir, em seus metadados, uma descrição referente ao tipo de cultura contido nos mapas, por exemplo: Aptidão climática ao plantio de abacaxi ou Aptidão climática ao plantio de algodão, o que identifica o tipo de cultura a que o dado se refere, facilitando a consulta ao banco.

Há propostas de padrões nos Estados Unidos e no Canadá (FGDC, 2001), com o objetivo de fornecer terminologia e definições comuns para conceitos relacionados aos metadados geográficos. A seguir descrevemos a proposta do FDGC, como exemplo de esquema de metadados.

O FGDC (*Federal Geographic Data Committee*) é um comitê entre agências para promover a coordenação do uso, troca e disseminação de dados espaciais nos EUA. O FGDC (2001) propõe um padrão que

especifica os elementos necessários para suportar os principais usos de metadados: ajudar a manter um investimento interno em dado espacial, pelas organizações; prover informação sobre o domínio de dados de uma organização; prover informação para processar e interpretar dados transferidos de uma fonte externa.

O padrão estabelece um conjunto comum de terminologia e definições para a documentação do dado geográfico, incluindo elementos para os seguintes tópicos: identificação, qualidade do dado, organização espacial do dado, referência espacial, informação sobre entidade e atributo, distribuição e referência do metadado (NSDI, 1997).

O padrão permite ao usuário saber: qual dado está disponível, se o dado atende suas necessidades específicas, onde achar o dado e como acessar o dado. Muitos dados estão disponíveis com este formato de metadados nos EUA onde, estados, governos locais ou do setor privado são incentivados a adotar o padrão para documentar seus dados.

O FGDC também patrocina a criação de uma *Clearinghouse* (*National Geospatial Data Clearinghouse*) um *Website* que guia usuários ao melhor dado espacial para seus projetos por meio de pesquisa a metadados. A intenção não é centralizar todos os dados geográficos em um local, mas prover *links* na Internet para distribuir *Websites* onde os dados são produzidos e mantidos. Gerenciadores documentam e disponibilizam seus dados, de acordo com o padrão, para a *Clearinghouse*, assim usuários podem achar facilmente uma informação, o que promove interoperabilidade entre organizações.

Como sua ênfase é na disponibilidade da informação, o padrão FGDC não especifica a maneira pela qual a informação está organizada nem o processo de transferência. Com exceção da parte de entidades e atributos, que pode revelar parte do significado do dado, as demais partes não descrevem a semântica da informação espacial.

O grande problema da proposta do FGDC, e do uso de metadados em geral, é a excessiva ênfase em informações que descrevem o processo de produção dos dados. Com relação à sintaxe, o padrão limita-se a indicar qual o formato em que os dados estão disponíveis. No aspecto semântico, suas informações são muito limitadas, pois o FGDC não adota o "modelo padrão" de geoinformação (campos e objetos). Adicionalmente, o padrão

do FGDC reflete os compromissos inevitáveis do "projeto de comitê", pois requer uma excessiva quantidade de informações (de aplicação questionável), com dezenas de formulários.

Em resumo, a substancial burocracia envolvida em adotar o padrão FGDC não se traduz em benefícios proporcionais. Estes fatos talvez expliquem porque sua adoção ainda está limitada e porque o consórcio OpenGIS propõe seu próprio formato para metadados.

9.2.4 Uso de ontologias

O grande fator limitante de conversão de dados são as diferenças de entendimento entre comunidades de usuários distintas. Diferentes visões da realidade geográfica sempre existirão por pessoas com culturas diferentes, pois a própria natureza é complexa e leva a percepções distintas. Neste caso seria interessante conviver com estas diferentes formas de conhecimento sobre a realidade e tentar criar mecanismos para implementar e combinar diferentes visões, ou seja, representar o conhecimento geográfico no computador buscando interoperabilidade pela equivalência semântica dos conceitos entre sistemas distintos. Neste sentido, são propostos trabalhos relacionados a Ontologias e seu uso para interoperabilidade e concepção de SIGs baseados em Ontologias (Fonseca *et al.*, 2000).

A Ontologia é uma disciplina filosófica que vem desde o estudo feito por Aristóteles sobre as categorias e a metafísica, e pode ser definida como o estudo do Ser e de suas propriedades. Para a comunidade de Inteligência Artificial, ontologias são teorias que especificam um vocabulário relativo a um certo domínio (Fonseca *et al.*, 2000) e descrevem uma dada realidade usando o conjunto de premissas de acordo com o sentido intencional das palavras deste vocabulário.

O uso de ontologias no desenvolvimento e uso de sistemas de informação leva ao que chamamos de Sistemas de Informação baseados em ontologias (Guarino, 1998). Fonseca *et al.*, (2000) propõe um SIG baseado em ontologias, composto por um editor de ontologias, por um servidor de ontologias, por ontologias especificadas formalmente e por classes derivadas de ontologias. A Figura 9.3 mostra o esquema de um SIG baseado em ontologias.

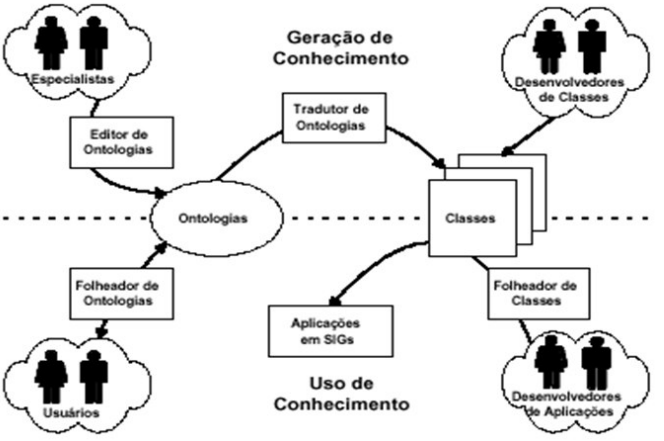


Figura 9.3 – Esquema de SIG baseado em ontologias (Fonseca *et al.,* 2000).

Na visão apresentada acima, os especialistas especificam formalmente seu conhecimento em ontologias, que são administradas por um servidor de ontologias. As ontologias são traduzidas para componentes de software por técnicas de orientação a objetos, constituindo um conjunto de classes, gerenciadas por desenvolvedores de classes, que formam uma estrutura hierárquica representando o mundo geográfico (Fonseca *et al.,* 2000). Desenvolvedores de aplicações utilizam o conhecimento traduzido em componentes de software (classes) para criar SIGs que também podem ser usados para troca de informações. O servidor permite o folheamento de ontologias e comunica-se com SIGs por meio de mediadores responsáveis por extrair informações destes e criar instâncias das classes que vão conter tal informação e o conhecimento extraído das ontologias.

Segundo Fonseca *et al.* (2000) a interoperabilidade semântica poderia ser resolvida através do uso de classes derivadas de ontologias, onde toda a manipulação de informações seria feita baseada nas definições das entidades geográficas presentes nas ontologias.

A interoperabilidade plena requer não só uma equivalência sintática entre as entidades representadas pelos sistemas, mas inclui também a equivalência de conceitos e significados destas entidades. Por exemplo,

duas comunidades de informação podem utilizar nomes diferentes para o mesmo conceito (como no caso de "rio" e "curso de água"). Ou ainda, um único conceito para uma comunidade (i.e., "rio") pode ser expresso com níveis maiores de detalhe por outra (i.e., "rios perenes", "rios temporários", "riachos"). Neste sentido, é necessário que os formatos de intercâmbio de dados disponham de um mecanismo que suporte o conceito de Ontologias e comunidades de informação geográfica. Com isto, interpretações diferentes de uma mesma realidade geográfica possam ser identificadas e facilmente trocadas.

9.3 Estratégias para integração e interoperabilidade e exemplos

9.3.1 Estratégias para integração e interoperabilidade

As estratégias para tratar dos problemas de integração e interoperabilidade entre sistemas de informação geográfica podem ser classificadas em quatro categorias principais (Gupta et al., 1999).

A abordagem mais simples consiste em definir catálogos de metadados e dicionários geográficos. Um *catálogo de metadados* armazena descrições de coleções de dados armazenadas em diversas fontes (Nebert, 2002), oferecendo serviços de localização, consulta e gerência de metadados, assim como serviços de solicitação de dados, que são repassados às fontes. Um catálogo, no entanto, tipicamente não armazena os dados em si.

Um *dicionário geográfico (gazetteer)* (Atkinson, 2002) define um vocabulário consistindo do identificador, localização e parte dos atributos dos geo-objetos de interesse. Um dicionário geográfico tipicamente cobre uma região bem definida, como um país, por exemplo.

A *estratégia de federação* (Sheth e Larson, 1990) assume que as fontes de dados mantêm autonomia, mas cooperaram para oferecer suporte a operações globais, que acessam dados em mais de uma fonte. Uma federação é *fracamente acoplada* se a responsabilidade de criar e manter a federação é dos usuários, não existindo controle centralizado. Já uma federação é *fortemente acoplada* quando existe controle centralizado para acesso às fontes de dados.

A *estratégia de armazém de dados (data warehouse)* (Miller e Han, 2001) consiste em: (1) extrair os dados das diversas fontes; (2) transformar os

dados extraídos para um modelo comum; (3) armazenar os dados transformados em um único repositório. Este enfoque é viável quando o número de fontes é pequeno e relativamente estável, não necessitando constantes atualizações nos dados extraídos. O processo de transformação pode se tornar particularmente difícil face à heterogeneidade dos dados, conforme já discutido na Seção 9.2.

A *estratégia de mediação* (Gupta et al., 1999) baseia-se em uma arquitetura de 3 níveis: camada de adaptação, com as fontes de dados acessadas através de adaptadores; camada de aplicação, com as aplicações que desejam acessar as fontes; camada de mediação, com um ou mais mediadores, que registra as fontes de dados conhecidas, e processa as consultas produzidas pelas aplicações. A estratégia de mediação será discutida em detalhe na Seção 9.5.

Por fim, a *estratégia híbrida* combina a estratégia de armazém de dados com mediação. Por exemplo, a arquitetura descrita em (Voisard e Schweppe, 1998) considera 4 camadas: a camada de aplicação, que recebe requisições das aplicações; a camada de serviços virtuais, que oferece um visão uniforme do sistema (um banco de dados virtual); a camada de serviços concretos, que gerencia as tarefas dos vários componentes; e a camada de serviços de sistema, que trata de serviços internos do sistema.

9.3.2 Exemplo de catálogo de metadados e dicionário geográfico

O projeto da *Alexandria Digital Library (ADL)* (Smith e Frew, 1995) (Smith, 1996) (Frew et al., 2000) exemplifica a construção combinada de um catálogo de metadados e de um dicionário geográfico, disponível na Web.

Os metadados para informação geo-referenciada combinam elementos dos padrões de metadados USMARC e FGDC (USMARC, 1976) (FGDC, 2001), já discutido na Seção 9.2.3. O primeiro é o padrão de metadados adotado para bibliotecas convencionais do governo dos EUA, enquanto que o segundo define metadados primariamente para catalogar objetos geográficos em formato digital, e é adotado pelas agências do governo dos EUA. A especificação completa dos metadados mantidos pela ADL contém cerca de 350 campos.

O dicionário geográfico da ADL contém nomes e características de acidentes geográficos oriundos do *US Geological Survey's Geographic Names Information System* e do *US Board of Geographic Names*. O primeiro lista o nome de cerca de 1,8 milhões de acidentes geográficos, organizados hierarquicamente em 15 categorias. Já o segundo contém os nomes de cerca de 4,5 milhões de acidentes geográficos, inclusive acidentes submarinos.

A arquitetura da ADL (ver Figura 9.4) segue um modelo de 3 níveis: *servidores ADL* gerenciam as coleções de dados; *mediadores ADL* implementam serviços sobre as coleções de dados; *clientes ADL* oferecem os serviços aos usuários finais.

Mais detalhadamente, um servidor ADL é responsável por manter uma coleção de metadados descrevendo os objetos catalogados e por implementar os mecanismos de consulta aos metadados, de acordo com os serviços definidos pela ADL. Uma entrada no catálogo pode incluir referências a representações digitais do objeto, disponíveis *online* ou *offline,* ou a representações não-digitais. Os servidores são autônomos, desde que ofereçam os serviços definidos pela ADL. Assim, uma instituição pode implementar um servidor ADL e publicar os seus dados através de um mediador ADL.

Um cliente ADL é responsável por oferecer os serviços ADL aos usuários finais, quer sejam usuários humanos ou agentes de software. Um cliente ADL deve manter o estado das sessões e suportar interações complexas com os usuários finais.

A peça central da arquitetura é o mediador ADL, que esconde a heterogeneidade dos servidores ADL através de uma coleção de serviços padronizados, oferecidos aos clientes ADL. Estes serviços são o cerne do projeto, pois expõem a funcionalidade pretendida para o catálogo e o dicionário geográfico da ADL.

Os serviços do mediador ADL não mantém estados de sessões. Ou seja, cada chamada a um serviço do mediador é tratada como uma transação por si só, e qualquer relacionamento entre duas chamadas deve ser codificado nos seus parâmetros. Esta decisão de projeto simplifica a implementação do mediador, mas torna mais difícil implementar consultas que são refinadas em sucessivas interações com o usuário.

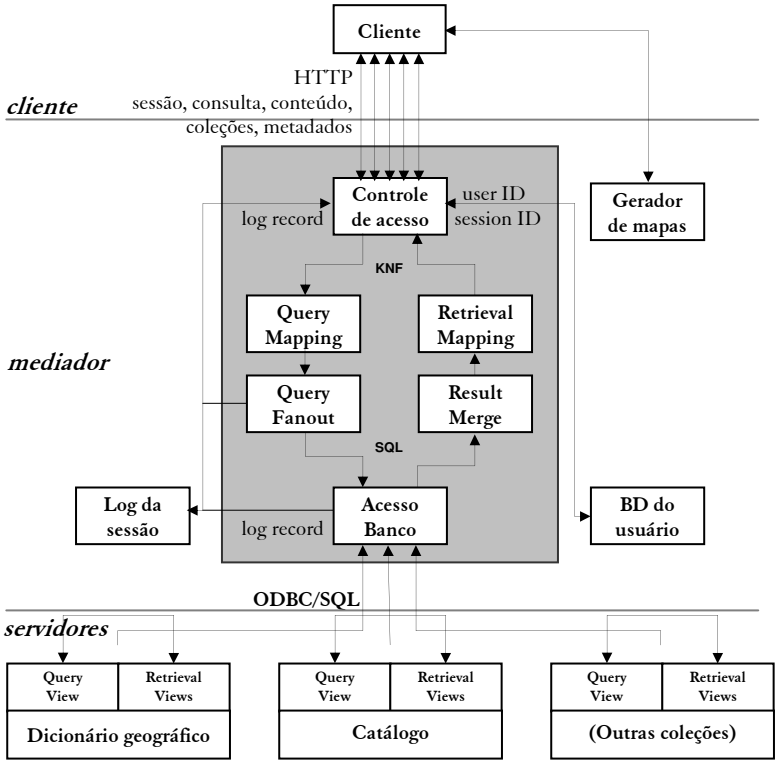


Figura 9.4 – Arquitetura da ADL (Frew et al., 2000).

9.3.3 Exemplo de armazém de dados geográficos

O *Projeto TerraServer* (Barclay, 2000) exemplifica a construção de um armazém de dados geográficos, combinado com um dicionário geográfico. O TerraServer representa o maior repositório público de imagens de sensoriamento remoto e mapas topográficos disponível na Web.

O TerraServer foi projetado para atender simultaneamente a milhares de acessos através da Web. Um usuário pode pesquisar o repositório de quatro formas diferentes:

1. Clicando em um mapa de baixa resolução da Terra, que indica onde há dados armazenados no repositório.
2. Indicando o nome de um local.
3. Indicando as coordenadas de interesse.
4. Selecionando o nome de um local de uma lista de locais bem conhecidos.

O repositório do TerraServer foi criado a partir de quatro fontes:

USGS Digital Ortho-Quadrangles (DOQ): imagens aéreas em cinza ou infravervelho na resolução de 1m. As imagens foram orto-retificadas de tal forma que 1 pixel corresponde a $1m^2$. Este acervo cobre aproximadamente 40% do território dos EUA, correspondendo a 3 milhões de quilômetros quadrados.

USGS Digital Raster Graphics (DRG): versões digitalizadas dos mapas topográficos do USGS. As imagens foram re-amostradas de tal forma que 1 pixel corresponda à potência de 2 mais próxima. Este acervo cobre todo o território dos EUA, correspondendo a 10 milhões de quilômetros quadrados.

Imagens do SPIN-2™: imagens em cinza na resolução de 1,56m, originárias de satélites militares russos. As imagens foram rotacionadas, com o norte para cima, mas não estão orto-retificadas; foram re-amostradas de tal forma que 1 pixel corresponde a $2m^2$. Este acervo cobre parte da Europa Ocidental, EUA e o Oriente, correspondendo a 1 milhão de quilômetros quadrados.

Encarta Shaded Relief: mapa em cores naturais do globo terrestre, indicando o relevo, obtido do CD-ROM da Enciclopédia Encarta. A imagem cobre continuamente o globo entre +80º e -80º de latitude, em uma resolução de aproximadamente $1km^2$ por pixel.

A arquitetura do TerraServer segue as três camadas usuais:

Cliente: um navegador normal que suporte HTTP 1.1 e HTML 3.2.

Aplicação: processa as requisições HTTP , repassando-as ao servidor de banco de dados.

Servidor de banco de dados: armazena os dados, servindo as requisições da aplicação.

A aplicação gera páginas dinamicamente, em HTML 3.2, e as envia ao navegador. Um usuário pode realizar operações de aproximação, afastamento e deslocamento nas imagens recebidas, sem necessidade de recursos especiais.

O banco de dados armazena mapas e imagens recortados em unidades menores, chamadas de *blocos (tiles)* nesta seção. Os blocos são agrupados logicamente em coleções contíguas, chamadas de *cenas*. Os blocos são indexados por tema, resolução, cena e localização.

O banco de dados mantém uma tabela para cada tema e resolução. Cada linha de cada uma destas tabelas contém os metadados de um bloco e um campo do tipo BLOB *(binary long object)*, armazenando o próprio bloco, no formato JPEG ou GIF.

Cada bloco é armazenado no banco de dados, redundantemente, em resoluções mais baixas, formando uma pirâmide de 7 níveis (ver Capítulo 13 para detalhes de armazenamento piramidal). O acervo do USGS/DOQ é compatível com resoluções de 1 a 64m; o acervo do USGS/DRG, de 2 a 128m; e o acervo do SPIN, de 1 a 64m.

O banco de dados também armazena um dicionário geográfico, permitindo ao usuário localizar cenas por nome de locais. O dicionário contém cerca de 1,5 milhões de nomes, incluindo sinônimos, e relaciona cada nome com o sistema de coordenadas utilizado pelo TerraServer. O dicionário contém ao todo 4 milhões de linhas e ocupa 3.3 GB.

O TerraServer entrou em operação em junho de 1998 e atualmente está entre os 1000 Web sites mais visitados. A média diária situa-se em: perto de 40 mil visitas; 3,6 milhões de acessos a blocos; 69GB transferidos.

9.3.4 Exemplo de mediador

A *Missão ao Planeta Terra (Mission to Planet Earth - MTPE)* é um programa da NASA para estudar processos ligados a mudanças climáticas, a partir de dados gerados por inúmeros satélites orbitando o planeta Terra. O *EOS Data and Information System (EOSDIS)* (Kobler, 1995) é o componente responsável por prover acesso fácil e rápido aos dados gerados no contexto do MTPE.

O EOSDIS é um sistema distribuído, organizado em três segmentos principais: o *Flight Operations Segment (FOS)* gerencia e controla os satélites e instrumentos do EOS; o *Communications and System Management Segment (CSMS)* fornece a infraestrutura de comunicação e gerência do sistema; e o *Science Data Processing Segment (SDPS)* trata do armazenamento, processamento e distribuição dos dados.

Em mais detalhe, o SDPS é o componente do EOS responsável por: aquisição de dados brutos; geração de dados derivados a partir dos dados brutos; arquivamento e distribuição dos dados derivados aos usuários. O SDPS está organizado em sete subsistemas principais:

Aquisição: recebe dados brutos e dispara processos para arquivá-los e processá-los.

Servidor de dados: fornece serviços de arquivamento físico e distribuição de dados, via FTP ou mídia removível.

Planejamento. fornece serviços para planejamento das atividades.

Processamento: executa os processos para geração de dados derivados.

Cliente: fornece uma interface para pesquisar e recuperar dados.

Interoperabilidade: fornece serviços para pesquisar e localizar serviços de dados.

Gerência de dados: fornece serviços para localizar e acessar dados.

O projeto Missão ao Planeta Terra gera vários Terabytes de dados diariamente e acumula um volume total de dados que chega a vários Petabytes, em formatos próprios, coletivamente chamados HDF-EOS. Portanto, o armazenamento e disseminação dos dados gerados representa um substancial desafio.

O projeto *NASA HDF-EOS Web GIS Software Suite (NWGISS)* (Di et al., 2001) visa exatamente tornar os dados gerados pelo EOSDIS disponíveis a outras aplicações que sigam as especificações do OGC (ver Capítulo 11 para um resumo dos padrões definidos pelo OGC e mencionados nesta seção). A arquitetura do NWGISS (ver Figura 9.5) consiste de três componentes principais: um servidor de mapas; um servidor de geo-campos *(coverage server)*; e um servidor de catálogo. O NWGISS inclui ainda um cliente OGC WCS.

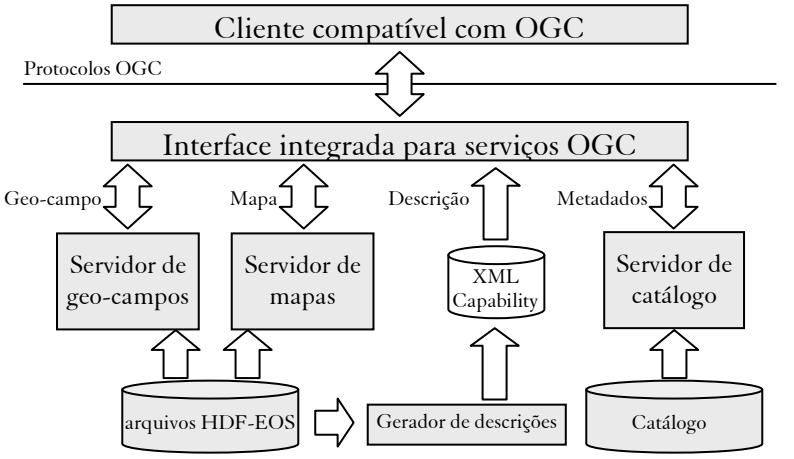


Figura 9.5 – Arquitetura do NWGISS.

O servidor de mapas do NWGISS implementa os serviços definidos no OGC WMS para todos os tipos de dados do formato HDF-EOS. Em particular, gera o geo-referenciamento do mapa em tempo de execução, se necessário. Da mesma forma, o servidor de geo-campos do NWGISS implementa os serviços definidos no OGC WCS para três formatos de dados, HDF-EOS, NITF e GeoTIFF. O servidor re-amostra, recorta e remonta geo-campos em tempo real, bem como transforma geo-campos de formato.

O cliente de geo-campos do NWGISS implementa a especificação do cliente OGC WCS. O cliente atua como um mediador para acessar geo-campos, nos formatos HDF-EOS, NITF e GeoTIFF, armazenados em servidores OGC WCS, não se limitando ao servidor OGC WCS implementado pelo NWGISS. O cliente permite acessar, visualizar, geo-retificar, re-projetar e reformatar geo-campos.

Como mediador, o cliente NWGISS é bastante limitado pois permite apenas selecionar geo-campos de mais de uma fonte, utilizando seus descritores, e visualizá-los em conjunto, entre outras operações.

9.4 Mapeamentos entre fontes de dados

9.4.1 Estratégias para definição de mapeamentos

Para interpretar e processar dados obtidos de diversas fontes, as aplicações devem ser capazes de tratar dados heterogêneos, tanto em formato e estrutura, quanto em interpretação ou significado. De fato, a heterogeneidade estrutural e semântica são problemas que bancos de dados distribuídos vem enfrentando desde longa data (Özsu e Valduriez, 1999).

Uma primeira estratégia para resolver o problema de tratar dados heterogêneos consiste em gerar mapeamentos entre pares de fontes de dados. Esta estratégia torna-se impraticável quando o número de fontes aumenta, ou quando não se conhece à priori as fontes disponíveis.

Uma segunda estratégia propõe uma descrição comunitária dos dados, chamada *esquema global ou federado*, *esquema mediado* ou *esquema de referência*, dependendo do enfoque adotado para atingir interoperabilidade (ver Seção 9.3.1). A esta descrição comunitária são então mapeadas as descrições das fontes de dados, chamadas de *esquemas locais* ou *esquemas exportados*, novamente dependendo do enfoque adotado. Esta estratégia simplifica o problema pois evita criar mapeamentos dois-a-dois entre os esquemas locais, mas ainda requer que as fontes sejam conhecidas à priori.

Uma terceira estratégia, proposta mais recentemente, adota ontologias para formalizar o esquema de referência e os esquemas locais. Utiliza ainda técnicas de alinhamento entre ontologias para simplificar a geração dos mapeamentos (Wache et al., 2001) (Uschold e Grüninger, 2001) (Mena et al. 2000).

Por fim, a definição dos mapeamentos entre as descrições locais e a descrição comunitária pode se beneficiar de uma análise dos metadados das fontes. Porém, segundo Haas & Carey (2003), a inexistência de metadados é um dos motivos para o fracasso de tentativas para atingir interoperabilidade entre fontes de dados. Além disso, metadados não necessariamente garantem a não ambigüidade dos termos, pois um mesmo termo pode ser usado em diferentes contextos, em diferentes línguas ou até mesmo de forma errônea. Uma possível solução seria

enfatizar o uso, uniforme e rigoroso, de metadados para completar a descrição dos conceitos pertinentes às fontes de dados. A própria descrição comunitária pode atuar como um vocabulário compartilhado, servindo de referencial à priori para a definição dos esquemas locais (Brauner, 2005).

9.4.2 Exemplo de definição de mapeamentos

Esta seção exemplifica questões compartilhadas pelas estratégias de federação, armazém de dados e mediação no que diz respeito à definição de mapeamentos.

Usaremos os termos *esquema de referência* para designar a descrição comunitária das fontes de dados e *esquema local* para descrever os dados visíveis em cada fonte de dados. Assumiremos que as descrições conterão definições de classes de objetos, denotando conjuntos, e de propriedades dos objetos, denotando funções mapeando objetos em objetos ou objetos em valores de dados.

Adotaremos a notação de teoria de conjuntos quando necessário. Desta forma, evitaremos escolher um particular modelo de dados para descrever os esquemas locais e o esquema de referência.

Um esquema local descreve as classes e propriedades dos dados que uma fonte deseja compartilhar, e o esquema de referência descreve as classes e propriedades dos dados oferecidos aos usuários como uma visão unificada das fontes. O esquema de referência contém ainda mapeamentos entre as classes e propriedades do esquema de referência e as classes e propriedades de cada fonte.

Há dois problemas que devem ser tratados na definição dos mapeamentos:

identificação de objetos: como definir uma forma universal de identificar os objetos nas várias fontes e, em particular, como identificar a ocorrência do mesmo objeto em fontes diferentes.

transformação das propriedades: como re-mapear os valores das propriedades oriundos de uma ou mais fontes para o esquema de referência, ou entre si.

Como exemplo, considere duas fontes de dados, F_A e F_B, descrevendo aeroportos. Suponha que o esquema local E_A da fonte F_A contenha (onde

C é o conjunto de caracteres adotado e $\Re$ denota o conjunto dos números reais):

(1) classes: Aeroporto
propriedades: Código: Aeroporto $\rightarrow C^3$
Cidade: Aeroporto $\rightarrow C^{50}$
Loc: Aeroporto $\rightarrow \Re^2$

e que o esquema local E_B da fonte F_B contenha:

(2) classes: Airport
propriedades: Code: Airport $\rightarrow C^3$
Name: Airport $\rightarrow C^{20}$
Coord: Airport $\rightarrow \Re^2$

Considere que o esquema de referência E_R contenha:

(3) classes: R-Aeroporto
propriedades: R-Código: R-Aeroporto $\rightarrow C^3$
R-Loc: R-Aeroporto $\rightarrow \Re^2$
R-Nome: R-Aeroporto $\rightarrow C^{20}$

Suponha que as propriedades Código, de E_A, e Code, de E_B, de fato armazenem os códigos universais dos aeroportos, com três caracteres. O mapeamento entre E_R, E_A e E_B pode então ser definido, por exemplo, como:

(4) R-Aeroporto =
$\{x \in \text{Aeroporto} / \exists y(y \in \text{Airport} \wedge \text{Código}(x) = \text{Code}(y))\}$

(5) $(\forall x \in \text{Aeroporto})(\text{R-Código}(x) = \text{Código}(x))$

(6) $(\forall x \in \text{Aeroporto})(\text{R- Loc}(x) = \text{Loc}(x))$

(7) $(\forall x \in \text{Aeroporto})(\text{R-Nome}(x) = n \Leftrightarrow$
$(\exists y \in \text{Airport})(\text{R-Cod}(x) = \text{Code}(y) \wedge \text{Name}(y) = n))$

Note que a sentença (4) indica que a classe R-Aeroporto de E_R é definida com base na classe Aeroporto de E_A. Assim, um aeroporto só estará disponível através do esquema de referência se e somente se for um aeroporto cadastrado na fonte F_A, por força da forma como o mapeamento foi definido (ver sentença (4)). Portanto, um aeroporto cadastrado na fonte F_B que não existir na fonte F_A não será visível através de E_R.

Assim, as sentenças (5) e (6) podem transferir diretamente as propriedades de aeroportos definidas em E_A para o esquema de referência E_R. Note que, arbitrariamente, E_R não contém a cidade do aeroporto. De fato, em geral, o esquema de referência não é a união dos esquemas locais.

A sentença (7) é mais complexa pois a fonte F_B pode conter aeroportos não cadastrados na fonte F_A. Portanto, é necessário verificar se o aeroporto existe na fonte F_A ao transferir a propriedade Name de E_B para o esquema de referência E_R.

Assuma agora que as propriedades Código e Code não representem os códigos universais dos aeroportos. Isto impediria identificar diretamente, via o código, quando dois aeroportos nas fontes F_A e F_B são o mesmo aeroporto. No entanto, se assumirmos que as propriedades Loc e Coord contém as coordenadas dos aeroportos no mesmo sistema de geo-referenciamento, então podemos substituir a sentença (7) por:

(8) $(\forall x \in \text{Aeroporto})(\text{R-Nome}(x) = n \Leftrightarrow$
$(\exists y \in \text{Airport})(\textit{Prox}(\text{Loc}(x), \text{Coord}(y)) \wedge \text{Name}(y) = n))$

onde *Prox* é uma relação de proximidade espacial que é verdadeira se e somente se dois pontos no espaço tiverem as mesmas coordenadas dentro de uma certa tolerância.

De fato, a adoção de um sistema universal de geo-referenciamento oferece também um sistema universal de identificação de objetos geográficos, ou pelo menos permite definir relacionamentos universais entre objetos geográficos.

Se assumirmos agora que as propriedades Loc e Coord contém as coordenadas dos aeroportos em sistemas de geo-referenciamento distintos, mas há suficiente informação nos metadados de Loc e Coord sobre os sistemas adotados, então devemos substituir a sentença (7) por:

(9) $(\forall x \in \text{Aeroporto})(\text{R-Nome}(x) = n \Leftrightarrow$
$(\exists y \in \text{Airport})(\textit{Prox}(\textit{Remap}(\text{Loc}(x)), \text{Coord}(y)) \wedge \text{Name}(y) = n))$

onde *Remap* é uma função que re-mapeia as coordenadas dadas por Loc para o sistema de geo-referenciamento adotado para Coord.

Este é um exemplo simples do problema de transformação das propriedades, no contexto de dados geográficos. A Seção 9.3.3 apresenta outros exemplos, quando discute a fase de transformação dos dados.

9.5 Construção de mediadores

9.5.1 Arquitetura de mediadores

Conforme adiantado na Seção 9.3.1, uma particular estratégia para integrar um conjunto de fontes de dados heterogêneas baseia-se na construção de um *sistema de mediação* com uma arquitetura em 3 camadas (ver Figura 9.6):

Camada de Aplicação: compreende as aplicações que desejam acessar as fontes.

Camada de Mediação: contém um ou mais mediadores fornecendo serviço de mediação para fontes de dados. Um mediador centraliza informações fornecidas por adaptadores, criando um *esquema mediado* das fontes de dados. O mediador também decompõe as consultas submetidas pelas aplicações em consultas a serem executadas pelos adaptadores, e reúne os resultados parciais para formar a resposta à consulta original.

Camada de Adaptação: contém os adaptadores responsáveis pelo acesso às fontes de dados. Cada adaptador esconde a heterogeneidade da fonte de dados, tornando o acesso à fonte transparente para o mediador. Para cada fonte de dados existe um adaptador que exporta algumas informações sobre a fonte, tais como: seu esquema, informações sobre seus dados e sobre seus recursos para processamento das consultas.

Ao participar de um sistema de mediação, uma fonte de dados, através do seu adaptador, deve ser capaz de expor um *esquema exportado*, descrevendo os dados que deseja tornar visíveis. Deve também oferecer serviços que permitam: (1) processar consultas sobre o esquema exportado; (2) transformar os resultados locais para os padrões definidos para intercâmbio de dados no sistema de mediação; e (3) aceitar temporariamente dados externos, convertendo-os para o formato local.

Por sua vez, o sistema de mediação deve ser capaz de expor um *esquema mediado* com a descrição comunitária dos dados, sobre o qual as aplicações definirão consultas ao sistema. O sistema deve ser capaz de executar consultas definidas sobre o esquema mediado, aplicando as transformações necessárias sobre os dados geográficos, além dos serviços usuais para definição e controle da execução de sub-consultas às fontes de

dados, combinação dos resultados e reestruturação dos dados convencionais. Esta seção aborda estes pontos em separado, seguindo (Gupta et al., 2000).

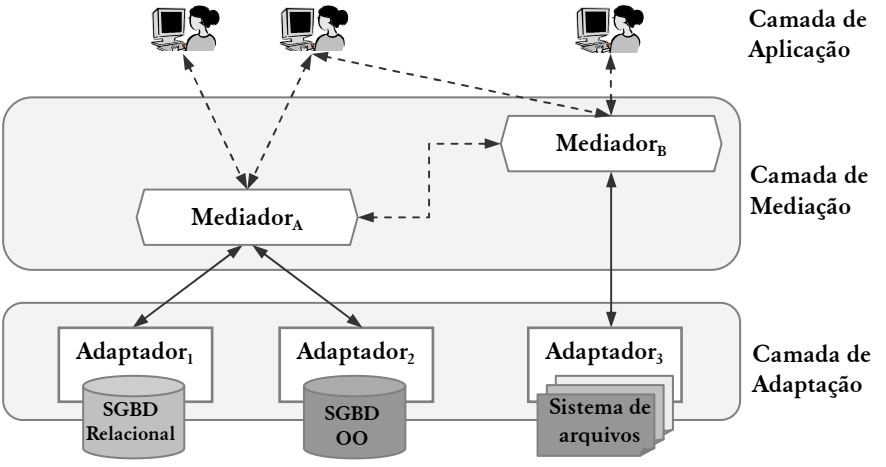


Figura 9.6 – Arquitetura mediador-adaptador (Gupta et al., 2000).

9.5.2 Serviços de adaptação

Acesso ao esquema exportado

Uma fonte de dados, através do seu adaptador, deve ser capaz de expor um esquema exportado, descrevendo os dados que deseja tornar visíveis, e o mapeamento entre este esquema e o esquema mediado. No caso de dados geográficos, o esquema exportado deve conter substancialmente mais informação para os atributos geográficos, conforme discutido a seguir.

Seja Q uma consulta sobre o esquema mediado. O mediador utiliza os esquemas exportados e os mapeamentos existentes no esquema mediado para selecionar fontes de dados e decompor Q em sub-consultas a serem enviadas às fontes selecionadas.

O processo de escolha das fontes deverá levar em consideração o custo de converter os dados geográficos armazenados em cada fonte para o formato adequado para processar a consulta. Diferentemente de dados convencionais, este custo poderá ser bastante alto e variar substancialmente de fonte para fonte, inclusive quanto à precisão do processo de conversão. Portanto, cada esquema exportado e o esquema mediado devem conter informação adicional sobre os atributos geográficos para subsidiar a decisão da escolha da fonte e da transformação necessária.

Em geral, cada atributo geográfico G deve estar associado a um *esquema de representação*, semelhante à noção de *measurement framework* definida em (Chrisman, 1997). O esquema de representação de G pode associar dois tipos de informação a G: (1) metadados, de forma semelhante aos esquemas de metadados discutidos na Seção 9.2.3 em outro contexto; (2) outros atributos cujos valores são necessários para interpretar o valor de G. O esquema de representação pode estar diretamente associado a G, ou ser herdado do objeto O ao qual G se aplica, ou da coleção de objetos a que pertence O. A Seção 9.5.4 apresenta exemplos do processamento de consultas em um mediador que utiliza esquemas de representação.

Processamento de consultas sobre o esquema exportado

Uma fonte de dados, através do seu adaptador, deve ser capaz de processar consultas sobre o esquema exportado, devolvendo os resultados no formato definido pelo mediador.

O formato utilizado pelo mediador para passar uma consulta para o adaptador deve ser especificado à priori, quando o sistema de mediação é criado. Por exemplo, o sistema de mediação pode adotar o padrão WFS, definido pela OGC (ver Capítulo 11), que especifica como as consultas devem ser passadas para as fontes de dados.

Transformação de dados

Uma fonte de dados, através do seu adaptador, deve ser capaz de aplicar transformações aos dados antes de enviá-los ao mediador.

A transformação mais básica consiste em converter os dados locais para o formato de intercâmbio de dados adotado pelo sistema de mediação. Por

exemplo, o sistema de mediação pode adotar o padrão GML, definido pela OGC (ver Capítulo 11), que especifica um formato de transporte para dados geográficos.

De forma geral, uma *transformação de dados* é qualquer função $f: D^n \rightarrow R$ em que os argumentos de f representam tanto dados armazenados na fonte, quanto parâmetros governando a transformação a ser aplicada aos dados.

Junto com a interface da transformação, o adaptador deve expor um modelo de custo, a ser utilizado pelo mediador ao montar o plano de processamento de uma consulta. Em geral, o modelo de custo captura a complexidade computacional da transformação, em função dos parâmetros de entrada e de uma estimativa do volume de dados recebidos como entrada e do volume de dados produzidos pela transformação.

9.5.3 Serviços de mediação

Acesso ao esquema mediado

O mediador deve ser capaz de expor o esquema mediado às aplicações, e fornecer meios para mantê-lo. Em particular, deve permitir o cadastramento de uma nova fonte de dados, com seu esquema exportado e mapeamento para o esquema mediado.

As questões levantadas pela exposição do esquema mediado às aplicações são semelhantes àquelas relativas à exposição do esquema exportado. Em particular, o esquema mediado deve associar cada atributo geográfico G a um esquema de representação.

Processamento de consultas sobre o esquema mediado

Um mediador deve ser capaz de processar consultas sobre o esquema mediado, devolvendo os resultados nos formatos solicitados pelas aplicações.

O mediador executa os seguintes passos ao processar uma consulta Q definida sobre o esquema mediado:

Seleção de fontes:

- baseando-se nos mapeamentos armazenados no esquema mediado e nos atributos utilizados na consulta, o mediador seleciona conjuntos de fontes capazes de produzir o resultado da consulta.

Otimização:

- para cada conjunto de fontes, o mediador utiliza os mapeamentos do esquema mediado para criar um plano, contendo sub-consultas de tal forma que cada uma possa ser inteiramente processada por uma fonte no conjunto, e que juntas produzam o resultado da consulta original.
- o mediador estima o custo de processamento de cada plano e escolhe o de menor custo estimado.

Execução:

- o mediador controla a execução do plano escolhido no passo de otimização.

Outras estratégias mais sofisticadas envolvendo transferências temporárias de dados de uma fonte para outra, por exemplo, podem ser definidas de forma similar às estratégias de otimização de consultas utilizadas em um banco de dados distribuído (Özsu e Valduriez, 1999).

Serviços adicionais

Um mediador pode utilizar as informações sobre as fontes de dados que armazena para oferecer serviços adicionais às aplicações.

Um primeiro exemplo de serviço adicional seria prover uma interface abrangente para que os usuários ou aplicações possam explorar o esquema mediado, incluindo os metadados mantidos sobre as fontes de dados. Desta forma, o mediador atuaria de forma semelhante a um catálogo de metadados, conforme ilustrado na Seção 9.3.2 e discutido em detalhe (Brauner, 2005).

Um exemplo deste tipo de serviço adicional, designado por *processamento cooperativo de consultas*, consistiria em aplicar transformações nas consultas submetidas para: corrigir as consultas; resolver ambigüidades; gerar consultas alternativas, quando a consulta

submetida falha; fornecer explicações sobre as respostas; complementar as consultas fornecendo informação adicional à explicitamente solicitada.

9.5.4 Exemplos de processamento de consultas sobre o esquema mediado

Considere duas fontes de dados, F_A e F_B, descrevendo aeroportos. Suponha que os esquemas sejam definidos da seguinte forma:

(10) Esquema exportado por F_A:

classes:	Aeroporto		
propriedades:	Código:	Aeroporto →	C^3
	Cidade:	Aeroporto →	C^{50}
	Loc:	Aeroporto →	$\Re^2$
	Ruído:	Aeroporto →	$2^{\Re_2}$

onde

- Código é o código universal de 3 letras identificando os aeroportos nos diversos países;
- Ruído(x) associa um geo-campo vetorial a cada aeroporto x representando o nível de ruído no entorno do aeroporto, e armazenado como uma grade regular de amostras pontuais (indicado na função apenas como um conjunto de pontos no $\Re^2$, assumindo que o espaçamento da grade é o mesmo para todos os aeroportos);

(11) Esquema exportado por F_B:

classes:	Airport		
propriedades:	Code:	Airport →	C^3
	Name:	Airport →	C^{20}
	Coord:	Airport →	$\Re^2$
	Noise:	Airport →	$\wp(\Re^2)$

onde

- Code é um código próprio da fonte FB (ou seja, não é o código universal);
- Noise(x) associa um geo-campo vetorial a cada aeroporto x representando o nível de ruído no entorno do aeroporto, e

armazenado como curvas de nível (representadas como elementos do conjunto $\wp(\Re^2)$);

(12) Esquema mediado E_M :

classes: R-Aeroporto

propriedades:
R-Código: R-Aeroporto $\rightarrow C^3$
R-Loc: R-Aeroporto $\rightarrow \Re^2$
R-Nome: R-Aeroporto $\rightarrow C^{20}$
R-Ruído: R-Aeroporto $\rightarrow 2^{\Re 2} \times \wp(\Re)$

onde

- R-Ruído(x) = (g,c) se e somente se g for a grade regular com amostras de ruído do aeroporto x, dada pela fonte F_A, e c for a curva de nível do ruído do aeroporto x, se este for cadastrado na fonte F_B, ou for o conjunto vazio, em caso contrário:

Considere ainda os seguintes mapeamentos, idênticos aos da Seção 9.4.2:

(13) R-Aeroporto =
$\{x \in \text{Aeroporto} / \exists y(y \in \text{Airport} \wedge \text{Código}(x) = \text{Code}(y))\}$

(14) $(\forall x \in \text{Aeroporto})(\text{R-Código}(x) = \text{Código}(x))$

(15) $(\forall x \in \text{Aeroporto})(\text{R- Loc}(x) = \text{Loc}(x))$

(16) $(\forall x \in \text{Aeroporto})(\text{R-Nome}(x) = n \Leftrightarrow$
$(\exists y \in \text{Airport})(\textit{Prox}(\textit{Remap}(\text{Loc}(x)), \text{Coord}(y)) \wedge \text{Name}(y) = n))$

e um último mapeamento capturando o significado pretendido para R-Ruído:

(17) $(\forall x \in \text{Aeroporto})(\text{R-Ruído}(x) = (\text{Ruído}(x),c) \Leftrightarrow$
$((\exists y \in \text{Airport})(\textit{Prox}(\textit{Remap}(\text{Loc}(x)), \text{Coord}(y))) \wedge c = \text{Noise}(y)))\vee$
$(\neg\exists y \in \text{Airport})(c = \varnothing)))$

Considere a seguinte consulta:

Q_1 : *Qual o nome e a localização do aeroporto com código GIG?*

O mediador procederá então da seguinte forma:

Seleção de fontes:

- de (14), (15) e (16), o mediador necessita acessar as duas fontes.

Otimização:

- novamente de (14), (15) e (16), o mediador decompõe Q_1 em duas sub-consultas, Q_{1A} e Q_{1B}, a serem enviadas a F_A e F_B, respectivamente:

Q_{1A} : Selecione a localização p=Loc(x) do aeroporto x
tal que Código(x)=GIG

Q_{1B} : Selecione o nome n=Name(y) do aeroporto y
tal que *Prox*(p', (Coord(y)))

- considere um único plano:

P_1 : Envie Q_{1A} à fonte F_A;
Espere o resultado contendo a localização p do aeroporto;
Re-mapeie a localização p para o sistema adotado na fonte FB, gerando uma nova representação p' da localização do aeroporto.
Envie Q_{1B} à fonte F_B;
Espere o resultado contendo o nome n do aeroporto;
Devolva a resposta (n,p);

Execução:

- o mediador executa o plano P_1.

Esta breve descrição deixa em aberto um ponto importante. É necessário que o mediador tenha acesso ao esquema de representação de Loc e Coord para que possa re-mapear p do sistema de geo-referenciamento de F_A para o sistema de F_B. Da mesma forma, o adaptador da fonte F_B, ou a própria fonte F_B, deve ter acesso ao esquema de representação de Coord para poder computar apropriadamente o relacionamento *Prox*.

Caso este segundo problema não possa ser resolvido por F_B ou seu adaptador, o mediador deve gerar um plano alternativo, substituindo a qualificação "*Prox*(p', (Coord(y)))" por "*dist*(p', (Coord(y))<k", onde o mediador decide o valor de k a partir dos esquemas de geo-referenciamento de Loc e Coord.

Considere agora a seguinte consulta:

Q_2 : *Obtenha o nível de ruído, em curvas de nível, do aeroporto na localização p.*

O mediador procederá então da seguinte forma:

Seleção de fontes:

- de (17), o mediador pode acessar qualquer uma das duas fontes.

Otimização:

- novamente de (17), há pelo menos 4 planos possíveis.
- O primeiro plano acessa a fonte F_A através da consulta:

Q_{2A1} : Selecione o nível de ruído C=Ruído(x) do aeroporto x
tal que Loc(x)=p

O plano consiste de:

P_{2A1} : Envie Q_{2A1} à fonte F_A;
Espere o resultado contendo o nível de ruído C do aeroporto;
Transforme (no mediador) C para curvas de nível C';
Devolva a resposta C';

- O segundo plano acessa a fonte F_A através da consulta, que já inclui a transformação:

Q_{2A2} : Selecione o nível de ruído C=Ruído(x) do aeroporto x
tal que Loc(x)=p, re-mapeando-o para curvas de nível

O plano consiste de:

P_{2A2} : Envie Q_{2A2} à fonte F_A;
Espere o resultado contendo o nível de ruído C do aeroporto;
Devolva a resposta C;

- O terceiro plano acessa a fonte F_B através da consulta:

Q_{2A1} :Selecione o nível de ruído D=Noise(x) do aeroporto x
tal que Coord(x)=p'

O plano consiste de:

P_{2B1} : Re-mapeie p para p' no sistema adotado em F_B;
Envie Q_{2B1} à fonte F_B;
Espere o resultado com o nível de ruído C' do aeroporto;
Se existir, devolva C' como resposta;
Caso contrário, execute o plano P_{2A};

- O quarto plano acessa a fonte F_B através da consulta:

Q_{2A2} :Selecione o nível de ruído D=Noise(x) do aeroporto x
tal que Coord(x)=*Remap*(p)

O plano consiste de:

P_{2B2}: Envie Q_{2B2} à fonte F_B;
Espere o resultado com o nível de ruído C' do aeroporto;
Se existir, devolva C' como resposta;
Caso contrário, execute o plano P_{2A1} (ou o plano P_{2A2});

- o mediador deve estimar o custo de processar os planos e escolher um deles.

Execução:

- execute o plano escolhido no passo de otimização.

Novamente, a descrição deixa vários pontos em aberto. Por exemplo, os planos P_{2A1} e P_{2A2} diferem na capacidade do mediador e do adaptador de F_A transformarem o geo-campo armazenado em F_A de uma grade regular para curvas de nível. Dependendo da capacidade de cada um destes componentes, um dos dois planos prevalecerá.

Além disto, esta transformação pode ser muito mais cara do que tentar acessar diretamente a curva de nível armazenada em F_B, mesmo correndo o risco de não encontrar o aeroporto em F_B e ter que recorrer a F_A de qualquer maneira. Portanto, os planos P_{2B1} e P_{2B2} podem ser muito mais eficientes do que os primeiros dois planos.

Da mesma forma, os planos P_{2B1} e P_{2B2} diferem em qual componente realizará a conversão de p para o sistema de geo-referenciamento adotado por F_B.

Em qualquer um dos casos, novamente é necessário que o mediador tenha acesso aos esquemas de representação de Ruído e Noise para decidir qual dos planos é viável, ou de menor custo.

9.6 Leituras Suplementares

Recomenda-se inicialmente um estudo da série de padrões relativos a dados geográficos publicados pela *International Standards Organization* (ISO):

- ISO 19115 Geographic Information – Metadata standard
- ISO 19107:2003 Geographic Information – Spatial Schema

- ISO 19109 Geographic Information – Rules for Application Schema
- ISO 19110 Geographic Information – Methodology for Feature Cataloguing
- ISO 19111:2003Geographic Information – Spatial Referencing by Coordinates
- ISO 19112:2003 Geographic Information – Spatial Referencing by Geographic Identifier

Em particular, o ISO 19115 define um esquema de metadados para descrever informação geográfica e serviços. O modelo em UML deste padrão está disponível como o ISO TC211 Harmonized Model. Estes padrões podem ser encontrados sob forma de ontologias em Islam et. al. (2004).

Há várias ontologias bastante interessantes definindo vocabulários geográficos. Duas delas merecem destaque:

- *CYC Geographic Vocabulary*, disponível em: http://www.cyc.com/cyc-2-1/vocab/geography-vocab.html
- *Alexandria Digital Library Feature Type Thesaurus*, disponível em: http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm)

Descrições de várias arquiteturas e tecnologias para facilitar a integração e interoperabilidade entre fontes de dados geográficos podem ser encontradas na literatura.

Abordagens gerais para o problema de integração e interoperabilidade cobrem desde catálogos de dados geográficos a mediadores capazes de processar consultas a fontes distintas (Abel, 1998) (DeVogele et al., 1998) (Bishr, 1998) (Laurini, 1998) (Jacobsen e Voisard, 1998) (Bergmann et al., 2000) (Tanin et al., 2002). Enfoques baseados em XML ou em *Web services* podem ser encontrados em (Gupta et al., 1999) (Gupta et al., 2000) (Alameh, 2003) (Manpuria et al., 2003). Há inúmeras iniciativas baseadas nos padrões do OGC, como (Boucelma et al., 2002) (Essid et al., 2004), para citar algumas referências. Mais recentemente, várias abordagens para construção de mediadores baseadas em ontologias foram propostas na tentativa de explorar o maior poder de expressão das linguagens para

descrever ontologias e as tecnologias para alinhamento de ontologias (Mena et al., 2000) (Wiegand e Zhou, 2001)

As áreas de armazém de dados e mineração de dados espaciais merecem atenção especial pelo potencial das aplicações. Um ponto de partida é o livro texto de Miller e Han (2001).

Referências

ABEL, D. J.; OOI, B. C.; TAN, K.-L.; TAN, S. H. Towards integrated geographical information processing. **International Journal of Geographical Information Science**, v. 12, n.4, p. 353-371, 1998.

ALAMEH, N. Chaining Geographic Information Web Services. **IEEE Internet Computing archive**, v. 7, n.5, p. 22-29, 2003.

ATKINSON, R. F., J., 2002, **Gazetteer Service Profile of the Web Feature Service Implementation Specification**, Open Geoscience Consortium.

BARCLAY, T.; GRAY, J.; SLUTZ, D. Microsoft TerraServer: A Spatial Data Warehouse. **ACM SIGMOD Record**, v. 29, n.2, p. 307-318, 2000.

BERGMANN, A.; BREUNIG, M.; CREMERS, A. B.; SHUMILOV, S. Towards an Interoperable Open GIS. In: International Workshop on Emerging Technologies for Geo-Based Applications. Ascona, Switzerland, 2000. p. 283-296.

BISHR, Y. Overcoming the semantic and other barriers to GIS interoperability. **International Journal of Geographical Information Science**, v. 12, n.4, p. 299-314, 1998.

BOUCELMA, O.; LACROIX, Z. E. M. A WFS-Based Mediation System for GIS Interoperability. In: 10th ACM international Symposium on Advances in Geographic Information Systems. McLean, VA, USA, 2002. p. 23-28.

BRAUNER, D. F. **Uma Arquitetura para Catálogos de Objetos baseados em Ontologias**. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2005.Departamento de Informática., 2005.

DEVOGELE, T. P., C.; SPACCAPIETRA, S. On spatial database integration. **International Journal Geographical Information Science**, v. 12, n.4, p. 335-352, 1998.

DI, L.; YANG, W.; DENG, M.; DENG, M.; MCDONALD, K. The prototype NASA HDF-EOS Web GIS Software Suite (NWGISS). In: NASA Earth Science Technologies Conference. Greenbelt, Maryland, USA, 2001. p.

ESSID, M.; BOUCELMA, O.; COLONNA, F. M.; LASSOUED, Y. Query Processing in a Geographic Mediation System. In: 12th Annual ACM International Workshop on Geographic Information Systems. ACM Press New York, Washington DC, USA, 2004. p. 101-108.

FEDERAL GEOGRAPHIC DATA COMMITTEE (FGDC). **Content Standard for Digital Geospatial Metadata WorkBook**. Reston, VA. Disponível em: <http://www.fgdc.gov/metadata/constan.html>.

FONSECA, F.; EGENHOFER, M.; BORGES, K. Ontologias e Interoperabilidade Semântica entre SIGs. In: Workshop Brasileiro em Geoinformática (GeoInfo2000), 2., São Paulo, 2000. **Anais**. São José dos Campos: INPE, 2000. p. 45 - 52.

FREW, J.; FREESTON, M.; FREITAS, N.; HILL, L.; JANÉE, G.; LOVETTE, K.; NIDEFFER, R.; SMITH, T.; ZHENG, Q. The Alexandria Digital Library architecture. **International Journal on Digital Libraries**, v. 2, n.4, p. 259-268, 2000.

GARDELS, K. The Open GIS Approach to Distributed Geodata and Geoprocessing. In: Third International Conference/Workshop on Integrating GIS and Environmental Modeling. Santa Fe, NM, USA, 1996. p. 21-25.

GUARINO, N. **Formal ontology and information systems**. Amsterdam, Netherlands: IOS Press, 1998, p.3-15.

GUPTA, A.; MARCIANO, R.; ZASLAVSKY, I.; BARU, C. Integrating GIS and Imagery through XML-Based Information Mediation. In: International Workshop on Integrated Spatial Databases. Portland, ME, USA, 1999. p. 211.

GUPTA, A.; ZASLAVSKY, I.; MARCIANO, R. Generating Query Evaluation Plans within a Spatial Mediation Framework. In: 9th International Symposium on Spatial Data Handling. Beijing, China, 2000. p. 8a18-8a31.

HAAS, L.; CAREY, M. Will federated databases ever go anywhere? In: Lowell Database Research Self-Assessment Meeting. Lowell, Massachusetts, USA, 2003.

ISLAM, A.S.; BERMUDEZ, L.; BERAN, B; FELLAH, S.; PIASECKI, M, **Ontologies for ISO Geographic Information Standards**. http://loki.cae.drexel.edu/~wbs/ontology/2004/09/bug/iso-19115/index.htm

JACOBSEN, A. H.; VOISARD, A. CORBA – Based Open Geographic Information Systems. In: European Conference on Parallel and Distributed Systems. 1998. p.

KOBLER, B.; BERBERT, J.; CAULK, P.; HARIHARAN, P. C. Architecture and design of storage and data management for the NASA Earth observing system Data and Information System (EOSDIS). In: 14th IEEE Symposium on Mass Storage Systems. Monterey, California, USA, 1995. p. 65.

LAURINI, R. Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. **International Journal of Geographical Information Science**, v. 12, n.4, p. 373-402, 1998.

MANPURIA, V.; ZASLAVSKY, I.; BARU, C. Web Services for Accuracy-Based Spatial Query Rewriting in a Wrapper-Mediator System. In: Fourth

International Conference on Web Information Systems Engineering Workshops (WISEW'03). Rome, Italy, 2003. p. 63-71.

MAPINFO, **MapInfo Professional User's Guide - Appendix J: MapInfo Interchange Format.** [online] <www.mapinfo.com/community/free/library/interchange_file.pdf>. Mar. 2001.

MAPINFO, **MapInfo Professional**.GIS Software. Troy, New York, 2002.

MARC, 1976, A MARC Format, in OFFICE, L. O. C. M. D., ed., Washington, D.C., Library of Congress Information Systems Office.

MENA, E.; ILLARRAMENDI, A.; KASHYAP, V.; SHETH, A. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. **International Journal on Distributed And Parallel Databases (DAPD)**, v. 8, n.2, p. 223-272, 2000.

MILLER, J. H.; HAN, J. **Geographic Data Mining and Knowledge Discovery**. Bristol, PA, USA: Taylor & Francis, Inc., 2001.

NATIONAL SPATIAL DATA INFRASTRUCTURE (NSDI). **Geospatial Metadata.** Disponível em: <http://www.fgdc.gov/publications/documents/metadata/metafact.pdf>.

NEBERT, D., 2002, Catalog Services Specification, Version 1.1.1, OpenGIS© Implementation Specification, Open Geoscience Consortium.

ÖZSU, M. T.; VALDURIEZ, P. **Principles of distributed database systems**. Prentice-Hall, Inc., 1999.

SHETH, A.; LARSON, J. Federated database systems for managing distributed, heterogeneous and autonomous databases. **ACM Computing Surveys**, v. 22, n.3, p. 183-236, 1990.

SMITH, T. R. A Digital Library for Geographically Referenced Materials. **IEEE Computer**, v. 29, n.5, p. 54-60, 1996.

SMITH, T. R.; FREW, J. Alexandria Digital Library. **Communications of the ACM**, v. 38, n.4, p. 61-62, 1995.

SPRING - Sistema de Processamento de Informações Georreferenciadas. Versão 3.5. [S.I.]: Instituto Nacional de Pesquisas Espaciais (INPE), 2001. 1 CD-ROM.

TANIN, E.; BRABEC, F.; SAMET, H. Remote Access to Large Spatial Databases. In: 10th ACM International Symposium on Advances in Geographic Information Systems. McLean, Virginia, USA, 2002. p. 5-10.

THOMÉ, R. **Interoperabilidade em geoprocessamento: Conversão entre modelos conceituais de sistemas de informação geográfica e comparação**

com o padrão OpenGIS. 1998. 196 f. (INPE-7266-TDI/708). Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, 1998.

USCHOLD, M.; GRÜNINGER, M. **Ontologies: Principles, methods and applications**. Knowledge Engineering Review, v. 11, n.2, p. 93 -155, 2001.

USMARC, 1976, A MARC Format, in OFFICE, L. O. C. M. D., ed., Washington, D.C., Library of Congress Information Systems Office.

VOISARD, A.; SCHWEPPE, H. Abstraction and Decomposition in Interoperable GIS. **International Journal of Geographic Information Science**, v. 12, n.4, p. 315 - 333, 1998.

WACHE, H.; VÖGELE, T.; VISSER, U.; STUCKENSCHMIDT, H.; SCHUSTER, G.; NEUMANN, H.; HÜBNER, S. Ontology-Based Integration of Information – A Survey of Existing Approaches. In: IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, WA, USA, 2001. p. 108 -117.

WIEGAND, N.; ZHOU, N. Ontology-Based Geospatial XML Query System. In: 4th AGILE Conference on Geografical Information Science. 2001.