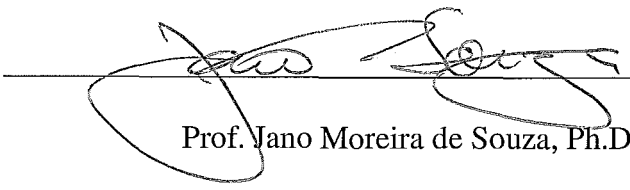
PUBLICAÇÃO DE DADOS AMBIENTAIS

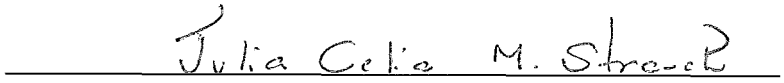
Gustavo da Rocha Barreto Pinto

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

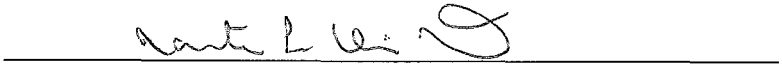
Aprovada por:



Prof. Jano Moreira de Souza, Ph.D.



Profª. Júlia Celia Mercedes Strauch, D.Sc.



Profª Marta Lima de Queirós Mattoso, D.Sc.

Profª Vânia Maria Ponte Vidal, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

JANEIRO DE 2003

PINTO, GUSTAVO DA ROCHA BARRETO

Publicação de Dados Ambientais

[Rio de Janeiro] 2003

XIII, 122 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2003)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Publicação de Bases de Dados Ambientais 2. Mediadores

3. Interoperabilidade de Bases de Dados Ambientais

I. COPPE/UFRJ II. Título (série)

A meus Pais

AGRADECIMENTOS

Ao Professor Jano Moreira de Souza, pela sua orientação, incentivo e oportunidades que me proporcionou na vida acadêmica e profissional.

À Professora Júlia Célia Mercedes Strauch, pela sua co-orientação sempre oportuna e adequada, suas valiosas sugestões e pela amizade.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro a este trabalho.

Aos membros da banca, Profª Marta Lima de Queirós Mattoso e Profª Vânia Maria Ponte Vidal pelas sugestões ao meu trabalho.

Aos meus colegas de mestrado, especialmente Nicolaas, Robson, Marcelo e André, pela amizade, companheirismo e conversas produtivas.

Aos colegas do SPeCS pelas reuniões em grupo, o apoio e incentivo durante o desenvolvimento e etapa final do trabalho.

À Patrícial Leal, secretária da linha de banco de dados, sempre disposta a ajudar.

Aos meus pais pela educação e apoio que me deram ao longo de toda a vida.

À minha querida Gláucia, pelo apoio, paciência e compreensão pela ausência, sem os quais não teria sido possível realizar o mestrado. Obrigado por tudo!

E a todos que contribuíram direta ou indiretamente para a conclusão deste trabalho.

Por fim, agradeço a Deus por tornar possível mais esta conquista.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

PUBLICAÇÃO DE DADOS AMBIENTAIS

Gustavo da Rocha Barreto Pinto

Janeiro/2003

Orientadores: Jano Moreira de Souza
Julia Celia Mercedes Strauch

Programa: Engenharia de Sistemas e Computação

Este trabalho apresenta a arquitetura de publicação de dados ambientais X-ARC – eXtensible Architecture. A arquitetura X-ARC é baseada no conceito de mediadores e é capaz de fornecer, de forma integrada, acesso aos repositórios de dados ambientais, distribuídos, heterogêneos e autônomos que se encontram espalhados através da Internet. A X-ARC visa facilitar o acesso dos usuários aos dados ambientais de forma a possibilitar seu compartilhamento, reutilização e interoperabilidade. Estes dados podem ser geográficos ou não-geográficos, estruturados, semi-estruturados ou não estruturados.

Para contemplar as diversidades destes dados, a X-ARC utiliza dois sistemas intermediários que auxiliam a execução das tarefas para prover os serviços de publicação e acesso aos dados: o Le Select , *midleware* do INRIA que publica dados estruturados e semi-estruturados; e o ArcIMS que publica dados georreferenciados.

Para suportar os diversos formatos existentes em cada repositório, a X-ARC utiliza a linguagem XML como padrão de intercâmbio de informação, provendo uma redução na heterogeneidade dos dados envolvidos e uma descrição da estrutura dos dados.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ENVIRONMENTAL DATA PUBLICATION

Gustavo da Rocha Barreto Pinto

January/2003

Advisors: Jano Moreira de Souza
Julia Celia Mercedes Strauch

Department: Systems and Computer Engineering

This work presents the environmental database publication architecture X-ARC - eXtensible Architecture. The X-ARC architecture is based on mediation concepts and is able to provide uniform data access to heterogeneous and distributed data sources spread over the Internet, maintaining their autonomy. X-ARC aims to provide to the users access to environmental data supporting data sharing, reuse and interoperability. These data can be geographical or non-geographical, structured, semi-structured or no structured.

In order to deal with data diversity, X-ARC uses two intermediaries systems to aid X-ARC tasks execution which provide publication services and data access services: Le Select middleware system developed at INRIA that publish structured and semi-structured data; and ArcIMS developed by ESRI that publish spatial data.

In order to provide access to several available formats in each data source, X-ARC uses XML language as a standard for information exchange, which provides data heterogeneity reduction and further data structure description.

SUMÁRIO

LISTA DE FIGURAS

1 INTRODUÇÃO

Com o surgimento da World Wide Web, a integração da informação evoluiu de uma arquitetura tradicional de múltiplas bases de dados para um novo *framework* capaz de manipular uma variedade de informações disponíveis em diversos formatos e estruturas. Este novo contexto da Internet criou novos temas de discussão para a integração de informações que são mais difíceis que os existentes para os sistemas de múltiplas bases de dados. Primeiramente, o número de fontes de dados podem ser enormes, tornando a resolução de conflitos e a integração de visões um problema difícil de ser resolvido. Segundo, o conjunto de fontes de dados é muito dinâmico, a inclusão ou a exclusão de uma fonte de dados deve ser realizado com o mínimo impacto sobre a integração das visões. Terceiro, as fontes de dados apresentam diferentes capacidades de processamento, variando de sistemas de gerenciamento de base de dados completos a simples arquivos de sistema. Quarto, as fontes de dados podem ser não-estruturadas ou semi-estruturadas não fornecendo desta forma nenhuma informação para a integração de visões. Por fim, a compreensão da semântica das aplicações requer uma especificação adequada dos metadados os quais são muito dependentes do domínio da aplicação (ÖZSU & VALDURIEZ, 1999).

Esta demanda também existe no âmbito da gestão de instituições que manipulam dados ambientais. Um exemplo é citado por STRAUCH (1998), que mostra em seu trabalho a necessidade de compartilhamento de dados ambientais e disserta sobre os meios utilizados para a troca destes dados e os esforços dos órgãos competentes, como *Open GIS Consortium* (OGC), para estabelecer uma especificação para a interoperabilidade dos dados.

1.1 Motivação

Na área ambiental, as atividades requerem o compartilhamento de dados e serviços. Em geral, estes dados são armazenados nas mais variadas formas de armazenamento (arquivos, banco de dados, etc), além de se apresentarem em diferentes tipos e formatos (texto, tabelas, objetos, etc), e possuírem diferentes capacidades de consulta. Além disso, encontram-se armazenados em diferentes servidores espalhados pela Internet com diferentes protocolos de acesso.

De acordo com TOMASIC & SIMON (1997), nesse domínio são encontrados dois tipos de usuários, a saber: usuários provedores e consumidores de dados. Usuários provedores buscam tornar seus dados públicos, enquanto que usuários consumidores necessitam localizar estes conjuntos de dados, de acordo com um determinado critério, a fim de visualizá-los, consultá-los, e eventualmente extrair alguma informação de seu interesse. Há ainda uma terceira denominação atribuída aos usuários que atuam em ambos papéis, usuários intermediários.

Duas dificuldades principais são encontradas quando buscamos desenvolver uma ferramenta de suporte ao compartilhamento de dados: (i) Heterogeneidade dos dados; e (ii) Distribuição dos dados.

Os diversos setores da área ambiental, tanto o público quanto o privado, vêm desenvolvendo bases de dados ambientais para apoiar os especialistas nas tomadas de decisão e análises. Entretanto, as análises e decisões envolvem a utilização de dados variados, de diferentes naturezas e provenientes de diversas fontes. Desta forma os dados manipulados pelos usuários nem sempre se encontram disponíveis na base de dados local existente. Surgindo então a necessidade de integração de informações das diversas bases de dados ambientais disponíveis e sua disponibilização aos usuários.

Atualmente, essa necessidade vem sendo suprida pela prática de intercâmbio de dados, de maneira informal, entre os diversos profissionais da área. Essa prática além de garantir a continuidade dos trabalhos, é responsável por atenuar o custo e a dificuldade de obtenção destes dados. Entretanto, devido à falta de um padrão de intercâmbio e à inúmera aplicabilidade destes dados, a heterogeneidade dos mesmos é alta além de possuírem diferentes semânticas.

Algumas iniciativas têm surgido para atender às necessidades destes profissionais, sendo que elas se apresentam geralmente na forma de arquiteturas ou padrões de dados. Entretanto, alguns aspectos como a semântica não foram totalmente resolvidos, o que torna necessária a interpretação dos dados pelo usuário, pois somente ele tem o conhecimento e a capacidade de abstrair o conjunto de dados necessário para sua necessidade específica.

1.2 Objetivos

Tendo em vista que a interoperabilidade entre as bases de dados ambientais é uma necessidade crescente entre as diversas instituições que trabalham com a informação, esta dissertação tem como objetivo o desenvolvimento de uma arquitetura, que possibilita o acesso, localização e interoperabilidade entre bases de dados ambientais heterogêneas e distribuídas que participam de uma cooperação entre os provedores e consumidores de dados.

O produto do esforço de pesquisa empreendido nesta dissertação é a arquitetura flexível e**X**tensible **Arc**hitecture (X-ARC). Na X-ARC, a interoperabilidade no nível semântico é alcançada utilizando uma camada de mediação, estruturada de forma hierárquica, baseada no conceito de Mediadores e

Tradutores. No nível de comunicação, a interoperabilidade é alcançada através da utilização dos padrões CORBA e RMI.

A arquitetura X-ARC encapsula dois sistemas para a execução de seus serviços e apoio à localização e acesso aos dados: o Le Select (LESELECT, 2002), *midleware* que manipula dados estruturados e semi-estruturados; e o ArcIMS (ARCIMS, 2001) que manipula dados georreferenciados. A arquitetura agrega novas funcionalidades sobre os sistemas, gerenciando informações sobre semântica, localização, estrutura, qualidade e segurança dos dados.

Para garantir a independência dos dados com relação a plataformas e formatos proprietários, é utilizada a linguagem *eXtensible Markup Language* (XML) como padrão de intercâmbio de dados. E através de um conjunto de metadados, específicos da arquitetura, coletados sobre os dados será possível ao usuário localizar e acessar dados ou fontes de dados de seu interesse.

Os dados acessados através da X-ARC são disponibilizados aos usuários em formato XML como uma coleção de dados que atendem aos critérios de pesquisa. Eles se apresentam com a estrutura e semântica de seus repositórios, cabe ao usuário a decisão de qual(is) conjunto(s) de dados utilizar e de que forma combiná-los .

Segundo ÖZSU & VALDURIEZ (1999), propostas visando a interoperabilidade de repositórios no contexto Web enfrentam o problema de manutenção da visão do mediador pois trata-se de um ambiente instável e dinâmico. Além disso, acrescenta-se o fato que a decisão de como unir, transformar e utilizar os dados é de responsabilidade dos usuários.

Desta forma, a arquitetura ao invés de integrar dados foi concebida para publicar dados, empregando o conceito de mediação e substituindo a complexidade

da manutenção da visão do mediador pelo uso de termos de domínio que agrupam os dados em coleções e metadados que descrevem os dados.

1.3 Contexto da Dissertação

Esta dissertação foi desenvolvida no contexto de três projetos de pesquisa do Programa de Engenharia de Sistemas e Computação da COPPE/UFRJ, a saber: Ambiente de Apoio à Decisão Cooperativa para Agricultura de Precisão (**CoAgri**), Sistema de Suporte à Decisão Espacial Colaborativo (**SPeCS**) e Ambiente AGROMET (**AGROMET**). O projeto SPeCS trata da especificação de um *framework* para suportar a decisão espacial colaborativa, enquanto que o AGROMET é uma instanciação do SPeCS na área de Agrometeorologia provendo um gerenciamento do conhecimento. Nesses projetos, o papel da arquitetura X-ARC é prover uma infra-estrutura, que possibilitará a integração de bases de dados ambientais heterogêneas e distribuídas envolvidas no processo de tomada de decisão e que participam de uma cooperação entre os detentores dos dados. Ela permitirá que os aspectos de heterogeneidade e interoperabilidade dos dados manipulados pelas áreas de aplicações dos projetos sejam abstraídos através do uso da X-ARC, permitindo uma concentração dos esforços nos objetivos principais dos projetos.

1.3.1 CoAgri

Trata-se de um projeto de cooperação entre o Programa de Engenharia de Sistemas e Computação da COPPE/UFRJ e o *Institut National de Reserche en Informatique et en Automatique* - INRIA. Este projeto visava desenvolver um

ambiente que apoiasse a decisão espacial colaborativa e auxiliasse a Agricultura de Precisão. Os objetivos do projeto evoluiram e foram englobados pelo projeto SPeCS.

1.3.2 SPeCS

Este projeto tem por objetivo fornecer um ambiente de trabalho cooperativo comum, flexível e fácil de usar onde os membros do grupo podem estar espalhados geograficamente em diversos ambientes heterogêneos e mesmo assim serem capazes de interagir durante um processo de tomada de decisão. Trata-se de um *framework* para suportar a decisão espacial colaborativa (MEDEIROS, STRAUCH, SOUZA et al., 2000, MEDEIROS, STRAUCH, SOUZA et al., 2001, MEDEIROS, 2002). Neste projeto são implementadas ações de pesquisa em:

I. Integração de Bases de Dados;

II. Ferramentas de Trabalho Cooperativo;

III. Ferramentas e Técnicas de Workflow;

IV. Modelagem Matemática para apoio à decisão; e

V. Acesso e manipulação de informações via Internet.

A Figura 1.1 apresenta as camadas que compõem a arquitetura SPeCS, onde observa-se que a camada de integração dos dados é representada pela arquitetura X-ARC com seus respectivos serviços.

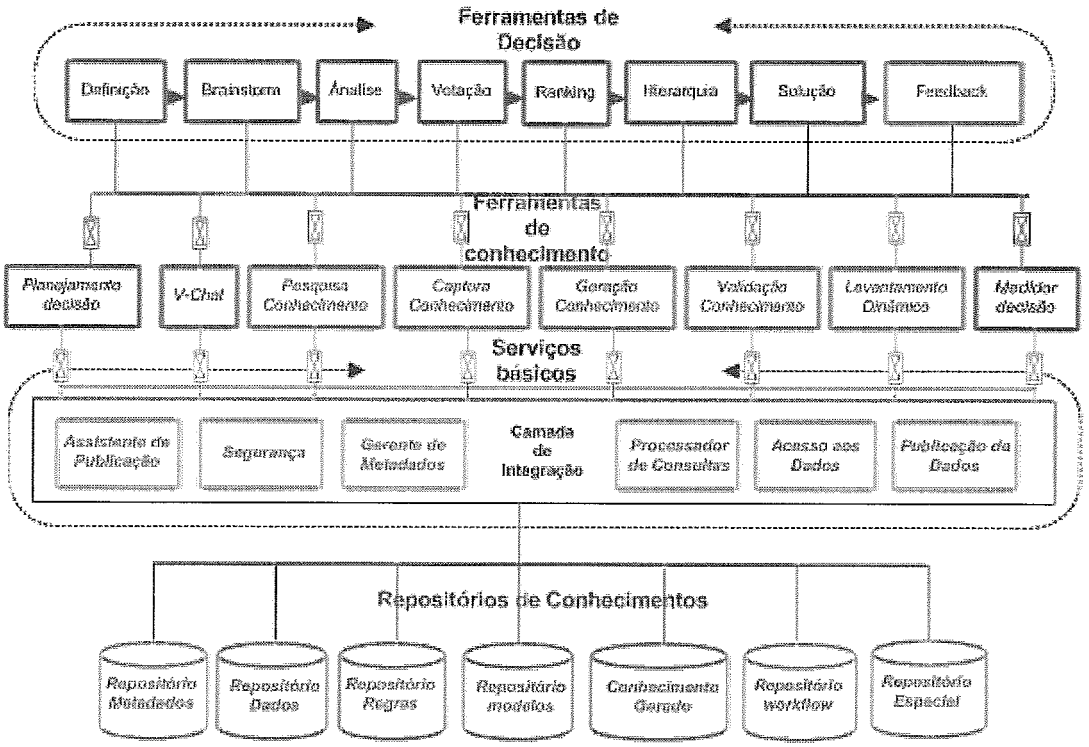


Figura 1.1 – Arquitetura SPeCS

1.3.3 AGROMET

Este projeto tem por objetivo fornecer aos seus usuários um ambiente para a gestão do conhecimento em Agrometeorologia, permitindo que o conhecimento obtido, gerado ou aplicado durante a execução das tarefas dos pesquisadores seja representado a fim de possibilitar sua utilização em atividades correlatas. A gestão do conhecimento neste projeto traduz-se em (SOUZA, STRAUCH, PINTO et al., 2002, PINTO, STRAUCH, SOUZA et al., 2002b):

- Implantar bases de dados e metadados agrometeorológicos que subsidiem os estudos dos usuários do ambiente;
- Desenvolver um subsistema para a integração de dados agrometeorológicos provenientes de diferentes instituições e em diversos formatos;

- Desenvolver um subsistema de apoio à decisão que empregue os modelos agrometeorológicos;
- Desenvolver um subsistema de *workflow* científico de modo a documentar e gerenciar as ações de pesquisas da Embrapa Solos voltadas ao monitoramento agrometeorológico;
- Desenvolver um subsistema de gestão do conhecimento científico para facilitar a integração de variáveis pedoambientais e uso do conhecimento interno;
- Desenvolver um subsistema para auxílio à escolha e/ou localização de novas estações agrometeorológicas para atender *workflows* das pesquisas da Embrapa Solos; e
- Desenvolver um subsistema para a gestão dos documentos utilizados ou gerados pela Embrapa Solos, na área de Agrometeorologia, facilitando a disponibilidade dessas informações a outras instituições parceiras da Embrapa Solos.

1.4 Organização do Trabalho

O presente trabalho está organizado em seis capítulos que são descritos a seguir. O capítulo 2 faz um breve relato sobre as características dos dados que permeiam o trabalho, apresenta as linguagens de publicação de dados na *Web*, com ênfase nas linguagens SGML e XML, a qual vem sendo utilizada como padrão de intercâmbio de dados. Ainda neste capítulo é feita uma dissertação sobre os metadados, a gerência de metadados e as principais propostas de arquitetura e padrões de metadados existentes.

O capítulo 3 apresenta uma revisão da literatura sobre integração de bases de dados com as abordagens de integração mais conhecidas como federação, mediadores e múltiplas bases de dados, e discute alguns trabalhos relacionados à arquitetura proposta.

No capítulo 4 é apresentada a especificação da arquitetura X-ARC. O modelo de mediação é apresentado e comparado com os trabalhos relacionados analisados no capítulo anterior. O modelo de metadados da arquitetura e seus componentes principais são discutidos e detalhados. Por fim, é apresentado o processamento de consultas definido para a arquitetura, assim como as ferramentas para a publicação e localização de dados disponíveis aos participantes da arquitetura.

Para exemplificar a utilização da X-ARC, um estudo de caso utilizando a arquitetura no escopo do ambiente AGROMET é apresentado no capítulo 5. Inicialmente, para compreender a contribuição da arquitetura para o projeto, é apresentado sucintamente o ambiente AGROMET e o papel da X-ARC no mesmo. Em seguida, é caracterizado o cenário do estudo de caso e as necessidades a serem consideradas. Ao final do capítulo, são apresentados os dados envolvidos no estudo, o protótipo desenvolvido para validar a arquitetura, as atividades envolvidas no processo de disponibilização dos dados e como os dados podem ser consultados, apresentando um consulta executada no protótipo e o resultado retornado.

O capítulo 6 finaliza este trabalho, apresentando os comentários conclusivos sobre as contribuições realizadas e indicando as perspectivas futuras que podem ser vislumbradas tendo como base a arquitetura proposta nesta tese.

2 REPRESENTAÇÃO DA INFORMAÇÃO

A fim de garantir um entendimento comum sobre o foco e o objetivo a que este trabalho pretende, é apresentado, a seguir, uma rápida definição e explicação sobre os dados ambientais que permeiam este trabalho. Durante a apresentação será possível apresentar os tipos em que se apresentam e se classificam.

Em seguida, é feita uma dissertação sobre os metadados, a gerência de metadados e as principais propostas de arquitetura e padrões de metadados existentes. Encerrando o capítulo, apresenta-se uma seção sobre linguagens de publicação de dados na web com ênfase nas linguagens de marcação SGML e XML e a arquitetura de metadados RDF.

2.1 Dados Ambientais

Neste trabalho, dados ambientais são caracterizados como:

Qualquer dado geográfico ou não-geográfico, estruturado, semi-estruturado ou não-estruturado que represente uma informação a respeito das entidades que compõem o meio-ambiente.

A partir desta definição, torna-se necessária a compreensão de dados geográficos, semi-estruturados e não-estruturados, que será apresentada nas próximas seções.

2.1.1 Dados Geográficos

Segundo STRAUCH (1998), dados geográficos são os dados armazenados nas bases de dados de Sistemas de Informação Geográfica (SIG), os quais representam

um conjunto de abstrações que descrevem elementos geográficos tais como, fenômenos, objetos, fatos físicos ou sociais do mundo real, sejam eles discretos ou contínuos, e suas relações com o meio. Eles são georreferenciados, ou seja, são referenciados a um sistema de coordenadas em relação à superfície terrestre.

Os dados geográficos são descritos em um domínio espacial, uma vez que possuem uma relação direta com a localização de um ponto ou porção da superfície terrestre. Eles são caracterizados por apresentarem dois componentes:

- O primeiro componente representa as propriedades gráficas dos dados geográficos, denominadas de atributos gráficos. Estes atributos descrevem a localização, a extensão e relacionamentos espaciais com outros objetos geográficos sobre uma representação, a qual irá constituir o mapa digital;
- O segundo componente, de atributos não gráficos, são caracterizados por dados alfanuméricos. Eles são armazenados como dados convencionais que descrevem propriedades qualitativas e quantitativas dos dados geográficos.

Os dados geográficos podem ainda ter atributos temporais e atributos descritivos. Os atributos temporais são decorrentes da característica temporal dos dados geográficos, uma vez que os fenômenos do mundo real podem variar sobre o tempo. Desta forma, é necessário também um sistema de referência temporal para definir uma época ou período de tempo. Os atributos descritivos, também denominados de pictóricos, contém uma descrição visual do objeto que auxilia a análise geográfica, como por exemplo, uma foto aérea, um desenho CAD ou uma imagem de satélite que contém uma cidade.

2.1.2 Dados Estruturados e Não-Estruturados

Dados estruturados são dados que apresentam uma estrutura fixa, definida e invariável na sua composição, ou seja, representam um tipo específico. Eles são manipulados pelos sistemas de bancos de dados tradicionais e por terem uma estrutura determinada, torna-se possível identificar um conjunto de restrições considerando tipos e relacionamentos. Devido a estas características, modelos podem ser estabelecidos para estes dados e esquemas podem ser definidos impondo restrições e validações sobre os dados manipulados.

Dados não-estruturados, pelo contrário, são dados totalmente ausentes de estrutura. Sobre eles é impossível determinar uma estrutura que possa servir para tipificá-los.

2.1.3 Dados Semi-estruturados

De acordo com ABITEBOUL (1997), dados semi-estruturados podem apresentar uma organização bastante heterogênea, que pode variar de um texto sem estrutura até um conjunto de registros bem formatados, ou seja, são dados que nem são totalmente desestruturados, nem totalmente tipados. Segundo o autor, as principais características dos dados semi-estruturados são:

- Estrutura irregular – as coleções destes dados consistem de elementos heterogêneos os quais podem estar incompletos ou podem trazer informações extras. Além disso, os elementos podem representar a mesma informação, porém serem representados por tipos diferentes, por exemplo: um endereço pode ser representado por uma seqüência de caracteres em um elemento e por uma tupla em outro.

- Estrutura Implícita – embora uma estrutura exista, ela se encontra implícita. Considera-se implícita mesmo que ela possa ser identificada por marcadores desde que: i) algum processamento seja necessário para obter a estrutura; ii) a correspondência entre a árvore "parseada" e a representação lógica dos dados não seja sempre imediata.
- Estrutura Parcial – estruturar completamente os dados é um objetivo, às vezes, difícil de alcançar, seja porque parte não possua estrutura ou seja porque exista apenas um esboço de estrutura sobre o dado. Desta forma, neste caso alguma parte da informação permanecerá desestruturada e armazenada de forma ineficiente sob o ponto de vista de banco de dados.
- Estrutura Indicativa versus Estrutura Restritiva – ao contrário das restrições de tipos existentes no bancos de dados tradicionais imposta pela tipificação forte, os dados semi-estruturados são orientados por uma estrutura indicativa dos tipos (*data guide*) a qual não impõe nenhuma restrição aos tipos dos dados apenas orienta sobre os tipos esperados. Dessa forma todo novo dado é aceito até mesmo quando uma alteração na estrutura indicativa se torna necessária e é custosa.
- Ausência de Esquema Pré-Definido – no contexto de dados semi-estruturados não existe a noção de um esquema pré-definido antes da inserção dos dados, mas sim um esquema após a inserção dos dados.

- Esquema Extenso – Como conseqüência da heterogeneidade, o esquema geralmente é muito extenso, diferente dos bancos relacionais nos quais o espaço ocupado pelo esquema é muito menor que o espaço ocupado pelos dados.
- Esquema Ignorado – tipicamente o esquema é ignorado para a execução de consultas sobre os dados semi-estruturados. Além disso, a inserção de dados pode desobedecer o esquema atual dos dados o qual é ignorado a fim de que se possa concluir a inserção.
- Evolução Rápida de Esquemas – No contexto de dados semi-estruturados, os esquemas são flexíveis e podem ser atualizados tão freqüentemente quanto aos dados, o que impõe grandes desafios para o seu gerenciamento.
- Tipos de Elementos de Dados Ecléticos – a estrutura de um elemento de dado pode variar dependendo do ponto de vista ou da etapa em que o processo de aquisição do dado se encontra. Um objeto pode ser primeiro um arquivo e depois ser uma coleção de objetos de referência (com estruturas complexas) tudo isso dependendo da etapa do processamento na qual este objeto se encontra. Por este motivo a noção de tipo é muito mais flexível.
- Esquema e Dados se misturam – A distinção entre esquema e dados para dados semi-estruturados não é muito clara e pode até não fazer muito sentido. Por exemplo, a informação do sexo de uma pessoa pode ser armazenado como dado em um repositório (valor booleano verdadeiro para masculino e falso para feminino)

ou como um tipo em outro repositório (o objeto pertence à classe Masculino ou Feminino)

Dados semi-estruturados apresentam uma representação estrutural irregular, não sendo nem completamente desestruturados nem estritamente tipados. Eles estão presentes de diversas maneiras e cada vez em um número maior de aplicações como banco de dados genéticos, banco de dados científicos, bibliotecas de programas e principalmente bibliotecas digitais, documentação *on-line*, comércio eletrônico. Torna-se fundamental compreender as questões envolvidas sobre estes dados a fim de desenvolver técnicas para o gerenciamento dos mesmos.

2.2 Metadados

Metadados são geralmente definidos como dados sobre dados e têm como objetivo facilitar o acesso, o gerenciamento e o compartilhamento de grandes conjuntos de dados estruturados e não-estruturados. Muitas pesquisas vêm sendo realizadas a fim de especificar metadados apropriados para aplicações com grande utilização de dados tais como: ciências da Terra, sistemas multimídia ou sistemas para mineração de dados. A maioria destes esforços resultam no desenvolvimento de conjuntos de metadados orientados à aplicação e, consequentemente, em padrões.

Metadados são utilizados extensivamente em sistemas e aplicações para garantir principalmente a eficiência no acesso, transferência, compartilhamento ou processo de grandes volumes de dados. A definição de um conjunto adequado de metadados é o primeiro passo para a implementação de sistemas e aplicações eficientes.

Durante os últimos anos, várias propostas foram feitas para atender as necessidades específicas de sistemas e aplicações. Entre elas podemos citar: FGDC

(FGDC, 1998) para sistemas de informação geográfica, Dublin Core (DUBLINCORE, 1999) para bibliotecas digitais e RDF (RDF, 1999) para a descrição de recursos na Web.

Entretanto, devido ao ambiente aberto criado pela Internet e a crescente necessidade de compartilhamento de informações surge a necessidade da interoperabilidade entre modelos de dados, sistemas e ferramentas. Esta mudança de visão implica em novas abordagens para o gerenciamento dos metadados as quais venham a permitir a integração, assim como a extensão das propostas já existentes.

Estabelecer maneiras de construir e relacionar recursos e suas descrições é a tarefa dos padrões e das arquiteturas de metadados. Os padrões de metadados são acordos que definem quais são os aspectos de um recurso que devem ser descritos. Em outras palavras, os padrões definem quais são os atributos de um recurso que interessam para sua descrição. Por exemplo, um padrão de metadados para bibliotecas apresentaria atributos como "Título", "Autor" ou "Data de Publicação". Cabe ressaltar que um padrão de metadados não é necessariamente definido em função de um domínio ou uma aplicação específica.

Além disso, a grande quantidade e diversidade de padrões de metadados que existem (e que existirão) gera a necessidade das arquiteturas de metadados. As arquiteturas de metadados são acordos que definem mecanismos através dos quais qualquer descrição de recurso indique qual o padrão de metadados (SAIF, FGDC, UDK, DUBLIN CORE) em que ela foi construída. As arquiteturas asseguram que uma informação seja corretamente interpretada, independente do padrão em que ela foi escrita e de quem a acesse. Assim sendo, ao manipular informações provenientes de padrões diferentes, não há a necessidade de unificar esses padrões antes de qualquer trabalho; a arquitetura trata de identificar a que padrão pertence cada

descrição. Logo, os recursos podem ser descritos seguindo não apenas um, mas diversos padrões, aproveitando o que eles têm de melhor em termos de semântica descritiva e garantindo a interoperabilidade entre eles.

Após essas considerações, pode-se afirmar que padrões e arquiteturas de metadados são complementares no processo de comunicação de informações. Como as aplicações têm diversas finalidades, existem diversos padrões de metadados, cada qual voltado a um determinado aspecto do domínio em que atua. Por esse motivo, as arquiteturas de metadados são essenciais para que se consiga atingir interoperabilidade entre informações descritas em diferentes padrões.

2.2.1 Gerência e Modelo de Metadados

Na literatura são encontrados trabalhos sobre gerência de metadados que propõem formas de melhor representar e modelar os metadados manipulados por arquiteturas específicas (LARRAONA, MOURA, MATTOSO, 1999) ou por aplicações em domínios variados.

KERHERVÉ & GERBÉ (1997) definem uma hierarquia de modelagem de metadados baseada em quatro níveis que deve ser implementada, caso tenha-se como objetivo principal um gerenciamento extensível dos metadados. Cada nível permite um grau de abstração variável, além de possibilitar a incorporação de modelos e meta-modelos para os metadados e a geração de metadados a partir da estrutura dos elementos de dados.

Na Figura 2.1, é possível observar os níveis de metadados e a hierarquia de abstração que esta abordagem permite. O **primeiro nível**, chamado **nível do dado**, corresponde às instâncias que são manipuladas pelos sistemas. Por exemplo, tuplas em um banco de dados relacional. O **segundo nível**, **nível de representação**, refere-

se a descrição dos conceitos usados para descrever os dados e metadados. Em um banco de dados corresponderia às instâncias existentes nos metadados mantidos pelo SGBD. O **terceiro nível, nível de meta-representação**, define o formalismo de representação utilizado no sistema. Ele descreve os conceitos aplicados na representação da informação no níveis inferiores. No caso de bancos de dados relacionais, trata-se dos conceitos aplicados na implementação do modelo relacional. A maioria dos sistemas possui apenas os três primeiros níveis. O **último nível**, denominado **nível de meta-meta representação**, permite uma representação homogênea dos outros níveis. Este nível permite uma interoperabilidade maior entre os modelos dos níveis inferiores. Exemplificando, seria como a partir dos conceitos do modelo relacional (nível 3) obtivéssemos os conceitos do modelo orientado a objetos (também nível 3) utilizando para isso um modelo geral.

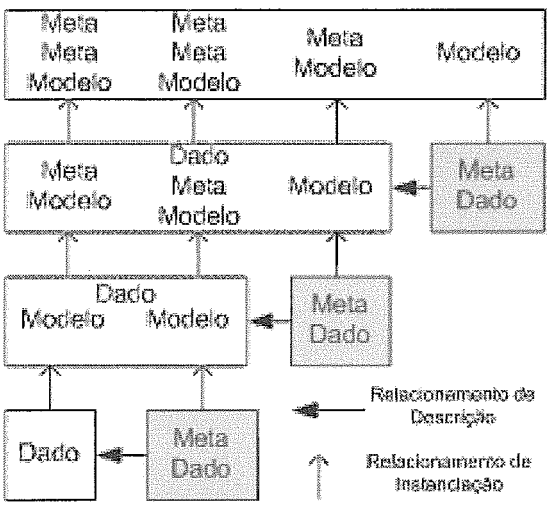


Figura 2.1 – Arquitetura de Múltiplos Níveis de Metadados

2.2.2 Padrões e Propostas de Metadados

Ao longo do tempo diversos padrões de metadados foram propostos com objetivos diversos, alguns foram concebidos para atender um domínio específico enquanto que outros, como por exemplo o UDK (GÜNTHER, LESSING,

SWOBODA, 1996), foram desenvolvidos para serem o mais abrangente possível. Dentre os padrões existentes, pode-se citar:

- SAIF – é um metamodelo que emprega o paradigma de orientação a objetos para definir uma hierarquia de classes que oferece estruturas de dados, independentes de *software,* para descrever dados geográficos em um nível mais abstrato do que os outros padrões de dados geográficos. Trata-se de um padrão canadense e foi projetado para modelar e promover a troca de dados espaço-temporais (SAIF, 1995, STRAUCH, 1998);
- FGDC – o padrão Federal Geographic Data Committee (FGDC), mais conhecido como "Content Standards for Digital Geospatial Metadata", especifica o conteúdo da informação de metadados para um conjunto de dados espaciais digitais (FGDC, 1998);
- UDK – o Catálogo de Dados Ambientais (GÜNTHER, LESSING, SWOBODA, 1996) é um projeto internacional que visa facilitar o acesso a dados ambientais heterogêneos e distribuídos. É um meta-sistema de informação, ou seja, contém dados sobre o formato e conteúdo dos dados disponíveis. Também documenta coleções de dados ambientais provenientes de diversas fontes (disponíveis *on-line* ou através de solicitação ao administrador dos dados);
- Dublin Core – o padrão Dublin Core (DUBLINCORE, 1999) consiste em um pequeno conjunto de atributos, apenas 15, que correspondem aos descritores mais comumente usados em qualquer padrão de metadados. O objetivo do Dublin Core é ser

um denominador comum entre todos os padrões de metadados de modo que a tradução e adaptação de um padrão para outro seja mais fácil;

- RDF – o *Resource Description Framework* (RDF, 2002, RDF, 1999) é uma arquitetura de metadados voltada para a descrição de recursos na Web. É fundamentada em um modelo de dados semântico muito simples, mas extremamente expressivo.

2.3 Linguagens de Publicação de Dados na Web

2.3.1 Linguagens de Marcação

Linguagem de Marcação é qualquer linguagem que permita acrescentar uma informação adicional ao texto de um documento para controlar a aparência do documento, tanto visualizado quanto impresso. Os marcadores são os responsáveis por delimitarem o início e fim de uma marcação. A marcação pode ser simples ou sofisticada, por exemplo:

- Simples:
  - Tipo e tamanho de fonte;
  - Negrito, sublinhado e itálico;
- Sofisticada:
  - Registro de índices;
  - Tabelas de conteúdo, etc;

Embora a maioria das marcações sirva para ajudar na aparência do documento, a **marcação generalizada** ou **descritiva** tem por objetivo indicar a importância estrutural de uma porção de texto dentro de um documento. Entre as linguagens existentes pode-se destacar SGML (SGML, 1986), HTML (HTML, 1999)

e XML (XML, 2002) como as mais conhecidas. A seguir, serão apresentadas as linguagens SGML e XML por serem as mais importantes considerando o aspecto de descrição e representação dos dados.

2.3.2 SGML

A *Standard General Markup Language* (SGML), ou Linguagem Padrão de Marcação Generalizada, é um sistema de marcação generalizada e foi publicado pela primeira vez em 1986 (LIGHT, 1999, SGML, 1986).

A SGML fornece um esquema de marcação simples, independente de plataforma e extremamente flexível. Além disso, não impõe qualquer conjunto específico de tipos de elementos, fornecendo maneiras de declarar tipos de elementos específicos. Pode-se definir a SGML como uma meta-linguagem, ou seja, uma linguagem para definir linguagens de marcação.

A característica marcante e principal da SGML é que o formato e a estrutura do documento são definidos através de elementos denominados marcadores. Pode-se identificar os marcadores em SGML através dos símbolos "<" e ">" que o acompanham, por exemplo, o marcador <SGML>.

A flexibilidade da SGML é alcançada através da definição do que é permitido ou não ocorrer em um documento. Esse conjunto de normas é denominado Definição de Tipo de Documento (DTD) e determina, basicamente, o seguinte:

- Os tipos de elementos permitidos no documento;
- As características de cada tipo de elemento, inclusive atributos permitidos e seu conteúdo;
- As notações que podem ser encontradas dentro do documento; e
- As entidades que podem ser encontradas dentro do documento.

Um documento SGML indica a que DTD ele segue, desta forma um documento é capaz de descrever sua própria estrutura. Embora seja flexível, a SGML é muito complexa e extensa, desta forma seu uso não foi disseminado.

2.3.3 XML

A *Extensible Markup Language* (XML) (BRAY, 1998, XML, 2000) é um nova linguagem, adotada pelo World Wide Web Consortium (W3C) (W3C, 2002), que complementa a HTML na troca de dados pela Web. Ambas as linguagens se baseiam na SGML.

A XML dá aos provedores de informação a liberdade de definir suas próprias estruturas para a informação que distribuem. Ela foi projetada para ser fácil de implementar e interoperável tanto com a SGML como com a HTML.

A principal diferença em relação à HTML é que a XML não descreve a apresentação de um documento, mas o seu conteúdo. Outra diferença importante é que a XML é extensível, isto é, permite a criação de novos marcadores, enquanto que o conjunto de marcadores da HTML é fixo. Os marcadores XML podem ser encadeados em qualquer nível de profundidade; embora a HTML permita, ela não reconhece o encadeamento de marcadores. A última, mas não menos importante característica da XML é que ela permite que seja definida uma gramática (DTD) para validação do documento, enquanto que a HTML, por não permitir criação de marcadores, prescinde de gramáticas outras que a do próprio HTML.

Para bem entender a XML, é importante considerar a sua ambivalência como linguagem para estruturação de documentos e como linguagem para definição de conteúdo. A XML, assim como também a SGML, foi originalmente criada como uma linguagem para definir estruturas de documentos, independente de sua apresentação.

Entretanto, a utilização da XML para definir conteúdo, seja a XML pura ou variações como o RDF, é bastante recente. Com isso, há algumas característica da XML, que são importantes na estruturação de documentos, mas que não são adequados para definição de conteúdo. Um exemplo é a questão da imposição de ordem dos elementos, que é fundamental para estruturar documentos, mas restritivo demais para representar conteúdo.

2.3.4 RDF

A *Resource Description Framework* (RDF) é uma arquitetura de metadados voltada para a descrição de recursos na Web. Ela é fundamentada em um modelo de dados semântico muito simples, proporcionando-lhe facilidade de uso; mas também é um modelo extremamente expressivo, de modo que ela pode abranger as mais diversas situações (RDF, 2002).

A RDF provê interoperabilidade entre aplicações que trocam dados inteligíveis por máquinas. A intenção do W3C é que a RDF se torne a principal arquitetura de metadados do mundo. Assim sendo, a RDF seria o principal meio de integração entre os diferentes padrões de metadados que já existem e os que existirão. Embora ambicioso, esse projeto tem grande chance de ter sucesso, pois a RDF possui diversas características que estimulam isso:

- A RDF possui um modelo de metadados que consegue ser ao mesmo tempo simples (poucos elementos) e fortemente expressivo (descreve recursos em qualquer nível de abstração e relaciona recursos de níveis diferentes);
- A RDF pode ser embutida em recursos de diversos formatos ou sintaxes, inclusive HTML e XML; e

- Possibilita a criação de esquemas RDF, ou seja, classes e relacionamentos no estilo orientado a objetos. Isso é especialmente útil para a aplicação de RDF em domínios específicos, de modo a facilitar sua adoção por uma comunidade específica;

O papel da RDF é importante para a troca de dados na Internet pois esse ambiente é heterogêneo demais para as soluções existentes (arquiteturas ou padrões) que são específicas aos domínios de aplicação. Desta forma, a RDF é uma arquitetura à altura da complexidade trazida pela Internet pois permite a interoperabilidade entre as diversas soluções existentes sem que as mesmas sejam abandonadas.

3 INTEGRAÇÃO DE BASE DE DADOS

Este capítulo tem por objetivo apresentar uma revisão da literatura sobre integração de base de dados, enfocando soluções como múltiplas bases de dados, federação e mediadores. Assim, inicialmente é apresentada uma taxonomia e a seguir é efetuada uma breve revisão dos trabalhos relacionados, visando avaliar as soluções existentes e evidenciar a necessidade da arquitetura proposta.

3.1 Estado da Arte

A década de 80 foi marcada pela proliferação dos sistemas gerenciadores de bancos de dados (SGBD) e a descentralização da informação nos setores das organizações. Embora tal fato tenha representado uma evolução na gerência das atividades dos setores, tornou-se evidente a necessidade de compartilhar informações entre as diversas unidades, a fim de permitir uma visão global da gerência do negócio para a tomada de decisões corporativas. Além da descentralização dos dados, estas organizações encontravam-se com uma diversidade tanto de sistemas (operacionais, SGBDs, redes, etc.) quanto de equipamento, resultando em variados cenários de integração. A partir deste momento iniciou-se a busca por soluções capazes de integrar os diversos SGBDs existentes nas organizações (SHETH & LARSON, 1990, SOUZA, 1986, WIEDERHOLD, 1992).

Existe uma grande variedade de soluções para o compartilhamento de dados. A literatura utiliza uma variedade de termos para descrevê-los, tais como: armazém de dados, bases de dados distribuídas, múltiplas bases de dados, bases de dados federadas, sistemas interoperáveis, mediadores e sistemas de consulta mediada.

Embora grande maioria da informação encontre-se armazenada em SGBDs, o surgimento e crescimento da Internet acrescentou a existência de repositórios de dados sem um sistema gerenciador, como por exemplo páginas html, planilhas eletrônicas, arquivos texto formatados, etc. Desta forma, as soluções existentes na área de banco de dados não podiam mais pressupor a existência de um SGBD, seja com capacidades idênticas ou não, como sendo os componentes locais envolvidos no processo de integração de dados.

Sendo assim, apresentamos a seguir uma taxonomia para as soluções de integração de bases de dados.

3.1.1 Uma Taxonomia das Arquiteturas de Integração de Bases de Dados

A maioria das taxonomias propostas para as arquiteturas de integração de bases de dados baseiam-se na análise das dimensões fundamentais dos sistemas distribuídos: autonomia, heterogeneidade e distribuição (ÖZSU & VALDURIEZ, 1999, SHETH & LARSON, 1990). Além destas dimensões, existem propostas que ainda utilizam outras dimensões, tais como: interoperabilidade (SHETH, 1998) e flexibilidade de evolução (BUSSE, KUTSCHE, LESER et al., 1999).

A taxonomia apresentada é baseada na análise da interoperabilidade, que tem sido o requisito básico para os sistemas de informação modernos das últimas duas décadas. SHETH (1998) divide os sistemas em três gerações, a saber:

- Geração I – caracterizada pela ênfase em alcançar a interoperabilidade entre os sistemas, buscando reduzir a heterogeneidade devido às diferenças entre SGBDs, tratando atualizações, consistência dos dados e mecanismos de transação.

Múltiplas bases de dados e sistemas de federação de bases de dados são os exemplos mais significativos desta geração;

- Geração II – caracterizada pela proliferação de uma variedade de dados (estruturados, semi-estruturados), pelo crescimento da Internet, uso de padrões e metadados. Federação de bases de dados que acessam componentes com protocolo de acesso a dados simples (não necessariamente SGBDs) e mediadores constituem os exemplos desta geração;
- Geração III – caracterizada por um maior nível de distribuição, autonomia e heterogeneidade entre os usuários, os repositórios de dados e a informação disponível. Além disso, esta geração disponibiliza serviços e operações e enfoca o problema no ambiente Web. Exemplos desta geração devem ser capazes de relacionar o conteúdo e representação das fontes de informação com entidades e conceitos do mundo real. As soluções desta geração devem utilizar metadados (orientado a domínio e baseado em conteúdo), contexto e ontologia para fornecer a integração dos dados. Mediadores e "*brokers*" são exemplos desta geração.

A seguir é apresentado uma síntese das principais soluções englobadas entre as três gerações da taxonomia.

3.1.2 Múltiplas bases de dados

Múltiplas Bases de Dados (MBD) são em alguns casos referenciados como bancos de dados heterogêneos e distribuídos. Esta foi a primeira abordagem que surgiu para prover o compartilhamento de dados entre SGBDs heterogêneos. É

baseada na completa integração dos múltiplos bancos de dados a fim de prover uma visão única (esquema global) dos dados.

MBD podem ser entendidos como um sistema distribuído que atua seja como um *front-end* para vários SGBDs locais ou como uma camada de um sistema global sobre os SGBDs locais. O sistema global provê funcionalidades completas de banco de dados e interage com os SGBDs locais através de suas interfaces com os usuários externos. Embora os nós locais possam manter algumas funções globais para conectar com o sistema global, os SGBDs locais são autônomos. O sistema global fornece meios, seja através de esquemas globais ou seja através de linguagens de múltiplas bases de dados, de resolver as diferenças na representação dos dados e funções das bases locais pois a mesma informação pode estar armazenada em diversos locais e em diferentes formatos. Devido a esta capacidade de resolução, os usuários globais podem acessar informações espalhadas em diversas fontes através de um pedido relativamente simples (HURSON, BRIGHT, PAKZAD, 1994QUOTE, ELMAGARMID, RUSINKIEWICZ, SHETH, 1999).

A principal vantagem desta abordagem é o fato dos usuários terem uma visão e acesso aos dados de forma consistente e uniforme, fornecendo uma transparência ao usuário a respeito da localização e heterogeneidade das bases de dados envolvidas. Embora esta abordagem seja eficiente no objetivo a que se propõe, o fato dos sistemas locais estarem fortemente acoplados ao sistema global resulta em desvantagens que precisam ser ressaltadas:

1. Baixa escalabilidade da solução, pois a inclusão ou retirada de fontes implica na confecção de um novo sistema global;
2. Há a necessidade de um especialista que conheça todos as fontes de dados envolvidas no processo para determinar o mapeamento das

funcionalidades do sistema global para os sistemas locais, pois automatizar o processo de integração não é possível;

3. Redução da autonomia dos componentes locais para auxiliar a resolução de conflitos semânticos, pois um prévio conhecimento semântico dos dados é necessário para o processo de integração global. Até pode ocorrer que o esquema de um componente local seja alterado para facilitar a integração;
4. Devido à existência de várias metodologias para alcançar a integração, pode ocorrer a perda de conhecimento semântico. Além disso, é difícil provar a corretude do esquema global.

Integrar através da abordagem de múltiplas bases de dados além de consumir muito tempo também tem grande possibilidade de falhar na geração do esquema global. Além disso, não é adequada para esquemas com mudanças dinâmicas e freqüentes, pois todo o processo tem que ser refeito.

3.1.3 Federação de bases de dados

O objetivo de uma arquitetura de Federação de Base de Dados (FBD) é remover a necessidade da integração estática de esquemas globais (ELMAGARMID, RUSINKIEWICZ, SHETH, 1999). A integração não precisa ser total como na abordagem anterior, mas dependente das necessidades dos usuários sobre os repositórios. Assim um FBD pode ser fracamente acoplado ou fortemente acoplado (SHETH & LARSON, 1990, BUSSE, KUTSCHE, LESER et al., 1999). A principal diferença entre o acoplamento fraco e o forte é que no fraco não existe um esquema global único enquanto que no forte existe um esquema global que integra os esquemas de todas as bases de dados da federação.

A FBD possui cinco níveis de esquema e foi estendida da arquitetura de três níveis de um SGBD tradicional para suportar as dimensões de distribuição, heterogeneidade e autonomia. Esta abordagem caracteriza-se por utilizar um modelo de dados próprio, denominado modelo de dados comum (MDC) e uma linguagem de comandos interna. Na Figura 3.1 pode-se observar a arquitetura genérica de uma FBD e os cinco níveis de esquemas: local, componente, exportado, federado e externo. Os esquemas local, exportado e componente encontram-se no MDC.

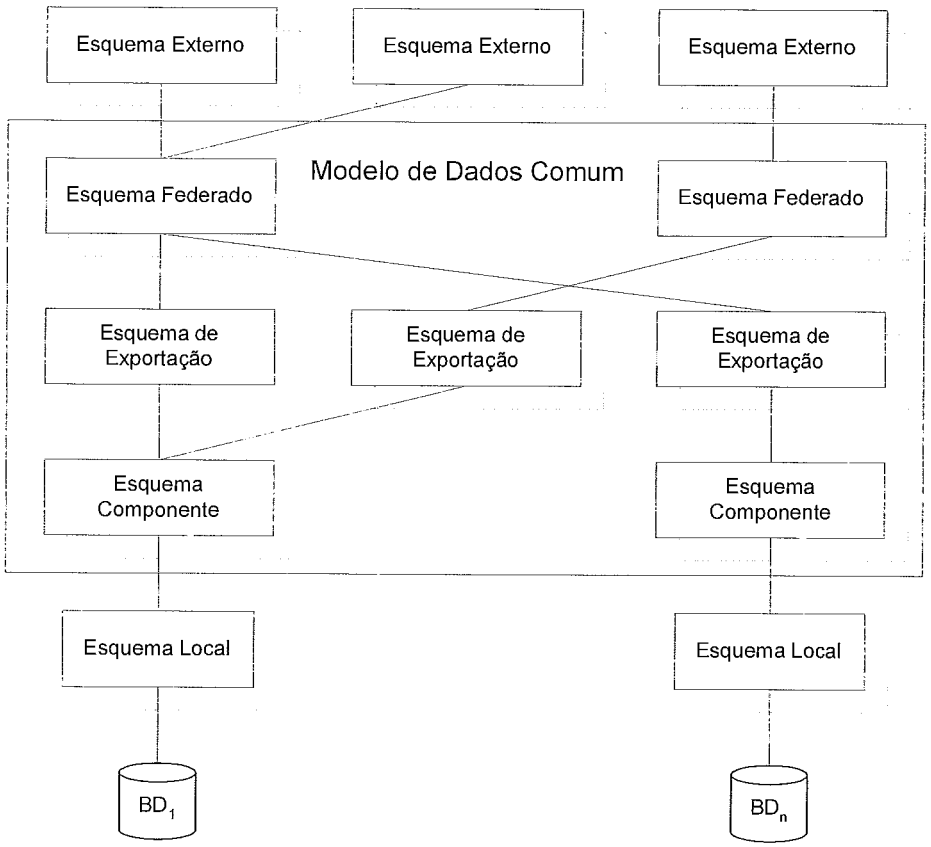


Figura 3.1 – Arquitetura genérica de Federação

A integração é alcançada através da conversão dos esquemas locais para o MDC (seja relacional ou orientado a objetos). Esta conversão tem por finalidade

homogeneizar a representação dos esquemas locais e facilitar a identificação dos conflitos entre eles.

O esquema global interno representa a visão integrada de todos os esquemas das bases de dados componentes da federação, expressas no MDC. Um exemplo da utilização desta abordagem para a integração de dados no contexto ambiental é a arquitetura MultiSIG proposta por STRAUCH (1998).

3.1.4 Mediadores

Segundo WIEDERHOLD (1992), um mediador é um componente de *software* que explora o conhecimento representado em um conjunto ou subconjunto de dados para gerar informações para aplicações residentes em uma camada superior. Cada mediador encapsula a representação de múltiplas fontes de informação, sendo responsável pela funcionalidade de acesso uniforme aos dados. Assim, os conflitos inerentes a integração de esquemas, tais como diferenças de nomes, diferenças de formato, diferenças estruturais e conflitos de valores são tratados por este componente (PIRES, 1997).

Os mediadores fornecem uma representação uniforme e flexível de fontes de dados arbitrárias, na qual o esquema para a visão integrada está disponível permitindo desta forma a execução de consultas e atualizações sobre este esquema. Existem duas abordagens empregadas na visão integrada do mediador: virtual e materializada. Na abordagem virtual, os dados se mantêm nas fontes locais e as consultas enviadas ao mediador são transformadas em subconsultas (nas linguagens de consulta de cada fonte) e enviadas para os repositórios locais. Cada resultado é então traduzido, filtrado e unido em um único resultado que é retornado para o usuário ou aplicação que originou a consulta inicial. Na abordagem materializada, ao contrário da virtual, a

informação de cada repositório é primeiramente extraída, filtrada, unida e armazenada em um repositório central. Então, a consulta enviada por um usuário pode ser avaliada diretamente no repositório, evitando que qualquer novo acesso seja feito às fontes de dados (VIDAL, LÓSCIO, SALGADO, 2001).

Uma classificação que segue esta abordagem é proposta em (DOMENIG & DITTRICH, 1999). Neste trabalho os autores discutem o estado da arte dos mediadores e dos sistemas baseados em mediadores, além de apresentar uma visão geral de alguns dos sistemas existentes na época em questão, trata-se de uma boa revisão para o assunto.

A Figura 3.2 apresenta uma arquitetura genérica de mediação em quatro camadas: aplicação, mediação, tradutores e fontes de informação, na qual, para cada fonte de dados respectivamente existe um tradutor.

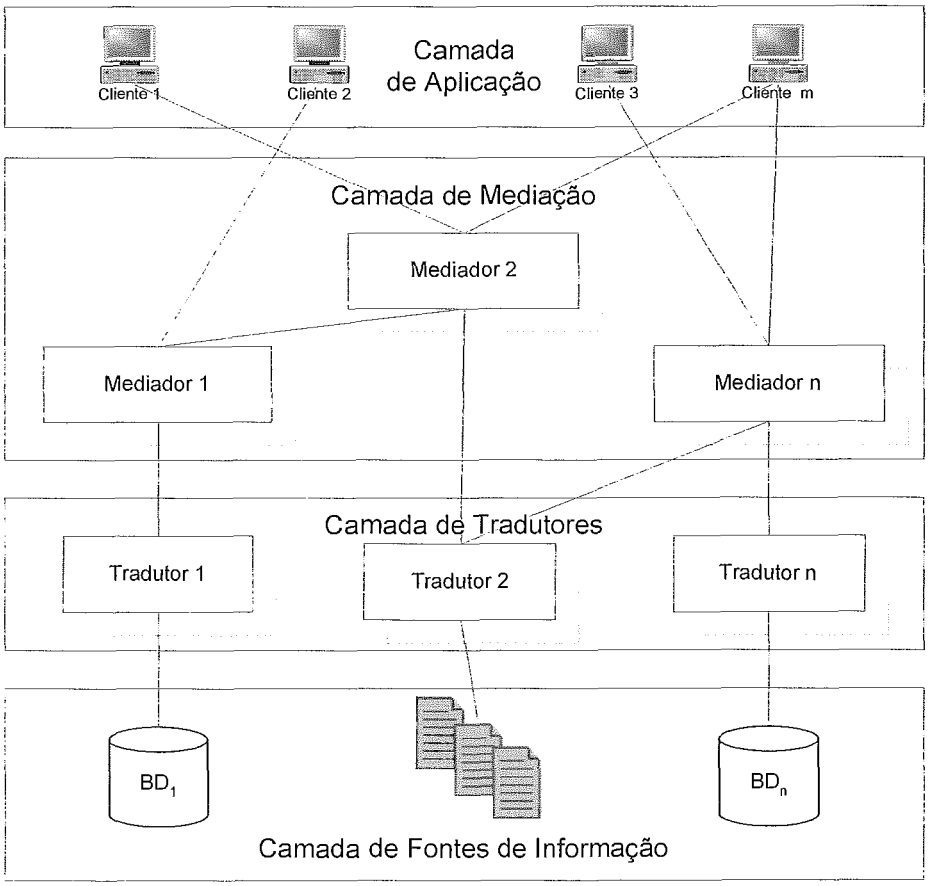


Figura 3.2 – Arquitetura genérica de Mediação

3.2 Trabalhos Relacionados

Entre as diversas soluções para o problema de integração de dados é apresentado a seguir alguns trabalhos que se relacionam com a arquitetura proposta por este trabalho.

3.2.1 Garlic

O Garlic é um sistema intermediário, baseado no conceito da arquitetura de mediadores, que provê uma visão integrada de diversas fontes de dados legadas, sem alterar onde e como o dado está armazenado. Foi desenvolvido pela IBM no *IBM Almaden Research Center*. Encapsula o modelo dos dados legados como objetos, os quais participam no planejamento da consulta. Ele provê interfaces padrão para a chamada a métodos e a execução da consulta. Trata o desafio da diversidade de padrões de como a informação das fontes de dados é descrita e acessada (CAREY et al., 1997).

O modelo de dados comum utilizado pelo sistema é o orientado a objetos que juntamente com a interface de programação são baseados no padrão estabelecido pelo *Object Database Management Group* (ODMG). O esquema unificado do Garlic é descrito por metadados globais mantidos pela arquitetura.

Devido a abordagem orientada a objetos, métodos podem ser associados a dados. Esta capacidade é explorada pelo Garlic provendo uma maneira natural e conveniente para modelar em fontes de dados não tradicionais a busca e a manipulação de dados.

No planejamento da consulta o tradutor e o mediador negociam dinamicamente a função que o tradutor irá desempenhar na execução da consulta,

pois é o único que realmente conhece as capacidades de pesquisa e acesso do repositório que ele encapsula. O processador de consultas desenvolve planos de execução para decompor, eficientemente, as consultas que envolvem múltiplos repositórios em subconsultas que os repositórios possam manipular. O executor da consulta controla a execução dos planos de consulta, montando os resultados dos repositórios e executando qualquer processamento adicional necessário para alcançar o resultado da consulta (ROTH & SCHWARZ, 1997, CAREY et al., 1997).

Uma visão genérica da arquitetura do Garlic é mostrada na Figura 3.3. Para cada repositório existe um tradutor. Além disso, o Garlic provê seu próprio repositório para objetos complexos, o qual pode ser criado pelos usuários para associar objetos de outros repositórios. Este repositório pode ser visto como um repositório de visões da arquitetura que interliga os dados dos repositórios existentes.

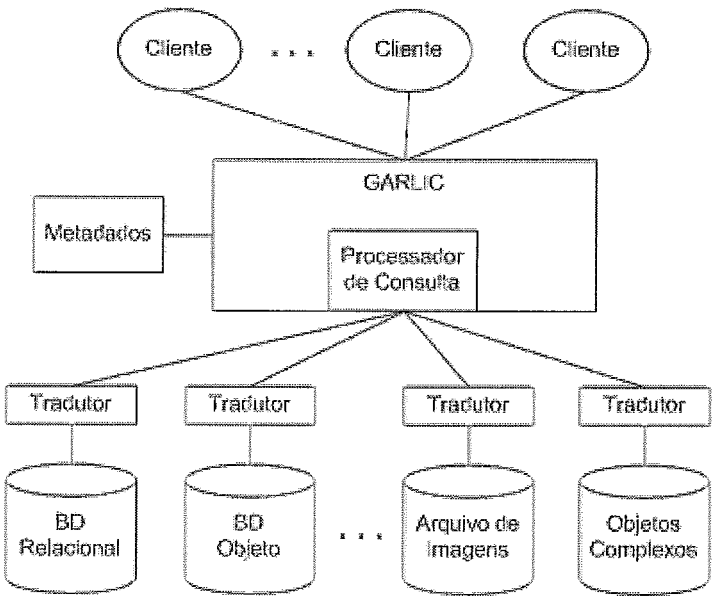


Figura 3.3 - Arquitetura do Garlic

A principal característica do Garlic é considerar no seu plano de execução a existência de repositórios de dados heterogêneos, com poder de consulta limitado ou mesmo inexistente e suprir essa limitação através do seu processador de consulta.

3.2.2 TSIMMIS

O projeto *The Stanford-IBM Manager of Multiple Information Sources* (TSIMMIS) é uma cooperação entre o IBM *Almaden Research Center* e a Universidade de Stanford que tem como objetivo fornecer ferramentas para a integração de fontes de dados heterogêneas estruturadas ou semi-estruturadas e verificação da consistência dos dados obtidos (CHAWATHE, GARCIA-MOLINA, HAMMER et al., 1994). No TSIMMIS, os mediadores são construídos sobre um determinado conjunto de fontes de dados, utilizando os tradutores que exportam objetos do modelo de dados comum do TSIMMIS denominado *Object Exchange Model* (OEM).

Este MDC é um modelo de objetos aninhados, autodescritivo e simples, semelhante à uma estrutura de grafo direcionado com arestas rotuladas. No modelo OEM, todas as entidades são objetos que podem ser atômicos ou complexos. Estes objetos são nós do grafo, no qual os objetos complexos têm arestas rotuladas com o tipo de relacionamento para seus subobjetos e os objetos atômicos contêm valores dentre os tipos atômicos (GOLDMAN, MCHUGH & WIDOM, 1999).

Uma fonte de dados é exportada como um conjunto de objetos OEM. Desta forma, os mediadores provêem visões OEM integradas dos dados encapsulados. Os mediadores são especificados através da linguagem *Mediator Specification Language* (MSL) que pode ser considerada uma linguagem para definição de visões, trata-se de uma linguagem orientada a objetos e baseada em lógica aplicada ao modelo OEM. A descrição do mediador é feita fornecendo regras lógicas que definem objetos OEM os quais o mediador disponibiliza em visões.

Os tradutores de cada fonte de dados são especificados com o auxílio da linguagem *Wrapper Specification Language* (WSL), que é uma extensão à MSL para

permitir a descrição do conteúdo das fontes e das capacidades de consulta. O objetivo é representar através de modelos (*templates*) as capacidades das fontes de dados. Cada modelo é associado a uma ação que gera os comandos para a fonte encapsulada.

O tradutor do TSIMMIS é mais pesado que os tradutores de outras arquiteturas de mediadores pois muitas das tarefas geralmente desempenhadas pelo mediador foram repassadas para o tradutor, como por exemplo a decomposição das consultas e a compensação pela ausência de capacidades de consulta.

As especificações do poder de consulta são expressas na linguagem *Query Description and Translation Language* (QDTL), que é uma gramática livre de contexto para a geração de consultas. Esta linguagem é tão complexa que uma ferramenta para a geração de tradutores, semelhante ao YACC, é fornecida para facilitar a tradução dos fragmentos da consulta para a linguagem de consulta do sistema.

Embora a utilização de especificações declarativas compactas para expressar o poder de consulta dos repositórios seja atraente, existem algumas desvantagens e problemas nessa abordagem. O primeiro e principal problema é a definição de uma linguagem para descrever todos as capacidades de um repositório, uma vez que é difícil capturar as restrições únicas associadas com cada repositório. Outro problema é o fato de que os modelos não são capazes de representar todas as possíveis consultas de serem executadas sobre o esquema da fonte de dados (KALINICHENKO, 2001, BUSSE, KUTSCHE, LESER et al., 1999).

A arquitetura do TSIMMIS possui ainda um componente que é responsável pelo gerenciamento das restrições sobre os dados integrados das diversas fontes de dados. Este componente fornece uma validação de restrições sobre os dados, que é menor se comparada com as restrições que um SGBD centralizado provê.

Uma desvantagem da abordagem de integração seguida pelo TSIMMIS é a necessidade de intervenção humana durante o processo, podendo em alguns casos, ocorrer da integração ser realizada manualmente pelo usuário, resultando possivelmente em erros subjetivos e de operação. Embora as dificuldades semânticas existentes nos repositórios geralmente não permitam a automatização total do processo de integração, a indispensável participação humana no processo do TSIMMIS é um fator crítico para o sucesso da integração.

O TSIMMIS não fornece um esquema do mediador, mas propaga todos os esquemas dos tradutores para o usuário, ficando sob responsabilidade do mesmo a decisão de que dado consultar e para onde enviar a consulta.

3.2.3 Le Select

O Le Select é o sucessor do DISCO (TOMASIC, RASCHID & VALDURIEZ, 1998), desenvolvido também pelo *Institut National de Reserche en Informatique et en Automatique* - INRIA.

O Le Select é um sistema intermediário que busca facilitar a publicação de fontes de informação distribuídas, heterogêneas e autônomas. O acesso aos dados e aos serviços de processamento é disponibilizado através da Internet ou Intranet. Os dados dos repositórios são mantidos na forma original e não precisam ser copiados para serem publicados (abordagem virtual) e o modelo de dados comum é o relacional.

Esta solução tem como objetivos principais:

i) permitir que proprietários de recursos (fontes de dados e programas) publiquem seus recursos para a comunidade, cabendo ao publicador escrever ou estender um tradutor adequado para o dado ou programa;

ii) fornecer uma interface uniforme dos recursos disponíveis para usuários em potencial e;

iii) permitir manipular os recursos disponíveis através de uma linguagem de alto nível.

Para cada recurso a ser publicado, é necessário o desenvolvimento ou configuração de um tradutor e o seu registro no servidor Le Select. Os esquemas dos dados publicados são conhecidos apenas pelos respectivos tradutores, não existe uma noção de catálogo global ou esquema integrado (TANAKA, VALDURIEZ, SIMON et al., 2001).

Em linhas gerais, a arquitetura do *Le Select* pode ser descrita através de quatro principais características:

1) O *Le Select* organiza a publicação de dados e serviços baseado no conceito de *site publicado,* ou seja, as fontes de dados e serviços tornam-se acessíveis para qualquer aplicação cliente que possa se conectar ao servidor Le Select;
2) Trabalha com padrões abertos e bem estabelecidos para interoperabilidade. A comunicação de rede entre seus componentes é realizada via protocolo CORBA e a interface local segue o padrão JDBC;
3) A publicação de dados ou serviços requer, apenas, o desenvolvimento de tradutores capazes de transformar os dados locais em relações ou mapear serviços locais dentro de consultas SQL;
4) O Le Select provê uma infra-estrutura que pode ser usada para suporte a metadados, pois permite a anexação de documentos no formato XML aos tradutores de serviços e dados, mas estes documentos não

precisam seguir nenhum padrão de metadados além de serem opcionais.

A separação entre clientes e servidores Le Select provê alta flexibilidade para camada de mediação, tanto para publicadores quanto para consumidores. Esta abordagem contrasta com prévios sistemas de mediação de informação tais como Garlic, DISCO e TSIMMIS, onde um mediador é associado com uma aplicação localizada em algum cliente que exporta um conjunto de visões definidas, isto é, um esquema global integrado.

Na arquitetura Le Select, mediadores são associados diretamente com publicadores ou clientes, não com aplicações. Como conseqüência, o Le Select não provê automaticamente uma transparência total para distribuição de dados, exigindo de seus clientes um conhecimento prévio da localização e semântica dos dados que desejam consultar.

3.2.4 MIX

O projeto *Mediation of Information using XML* (MIX) é uma colaboração entre o *UCSD Database Laboratory* e o grupo *Data-Intensive Computing Environments* (DICE). O objetivo do projeto é estudar, desenvolver, aplicar e avaliar sistemas de consulta sobre fontes de dados heterogêneas que utilizam XML (MIX, 2002, BARU, GUPTA, LUDASCHER et al., 1999). MIX baseia-se na arquitetura de mediadores (WIEDERHOLD, 1992) e emprega XML (XML, 2002) para representar de forma flexível e uniforme os dados existentes nos repositórios.

A abordagem do MIX não exige a conversão dos repositórios de dados convencionais para XML, entretanto pressupõe que uma visão lógica das fonte de

dados (seja bancos de dados, coleção de páginas html, ou até mesmo um sistema de dados legados) seja fornecida em XML.

Os repositórios são consultados utilizando uma linguagem de consulta XML própria, denominada *XMAS* (LUDASCHER, PAPAKONSTANTINOU & VELIKHOV, 1998). *XMAS* é uma linguagem funcional, que é influenciada pela *OQL* e possui um redução precisa para a álgebra orientada a tuplas da *XMAS*. Ela se diferencia da linguagem *XQuery* pois é orientada à tupla, o que permite o processamento de consultas de forma compatível com o processamento de consultas de bancos de dados relacionais, relacional aninhado e orientado a objetos.

No MIX, o resultado de qualquer consulta é um documento XML. Geralmente, o mediador do MIX recebe uma consulta, a decompõe em fragmentos (subconsultas) de acordo com as capacidades de consulta dos repositórios e envia os fragmentos para os repositórios apropriados. À medida que os resultados das subconsultas são retornados pelos repositórios para o mediador, ele integra os fragmentos em um único resultado e então retorna para o usuário.

Os tradutores são responsáveis por traduzir as consultas XMAS para consultas ou comandos que os repositórios encapsulados sejam capazes de compreender. Além disso eles convertem os resultados retornados pelos repositórios para o formato XML.

Cada repositório exporta o modelo dos dados que possui na forma de um DTD XML, que é utilizado como o esquema dos dados pelos componentes da arquitetura do mediador. Os dados são exportados pelos tradutores como documentos XML que atendem ao DTD especificado.

Ao contrário de outras abordagens, que baseiam-se em modelos semi-estruturados sem esquemas, o MIX utiliza XML em seu MDC e explora a estrutura da

informação fornecida pelos DTDs que descrevem os repositórios. Semelhantemente ao TSIMMIS, a avaliação da consulta é virtual.

3.2.5 Comparação dos Trabalhos Relacionados

Uma comparação entre os trabalhos relacionados é apresentado pela Tabela 3.1. Nesta tabela podemos observar que a maioria dos trabalhos baseiam-se na arquitetura de mediadores e que a abordagem utilizada é a virtual. Observa-se ainda que o diferencial entre os trabalhos é o modelo de dados comum adotado, o tipo de repositório de dados manipulado e a forma como os resultados são retornados.

Tabela 3.1 – Comparação entre os Trabalhos Relacionados

Características	**Trabalhos Relacionados**			
	Garlic	***TSIMMIS***	***Le Select***	***MIX***
Arquitetura de Integração	Mediador	Mediador	*Midleware*	Mediador
Autonomia	Alta	Alta	Alta	Alta
MDC	ODMG-93	OEM	Relacional	XML
Transparência	Linguagem	Linguagem, localização e esquema (visão MSL)	Linguagem	Linguagem
Atualização dos dados	Não	Não	Não	Não
Sentido de Integração	---	Top-down	---	Bottom-up
Tipos de dados	Estruturados	Estruturados e semi-estrutu-rados	Estruturados , semi-estruturados e programas	Estruturados e semi-estrutu-rados
Abordagem de	Virtual	Virtual	Virtual	Virtual

Mediação				
Esquema Global	Sim	Não	Não	Não

3.3 Proposta na Área Ambiental

Na área ambiental são encontradas algumas propostas tais como o Earth View (EARTH VIEW, 2002). Grande parte destas propostas visam fornecer acesso a dados ambientais heterogêneos e distribuídos, entretanto a maioria destas soluções são limitadas no aspecto da integração dos dados, fornecendo em suas propostas a possibilidade de visualização dos dados ou limitando seus escopos apenas aos dados espaciais. Propostas como o Earth View encontram-se na abordagem que permite a visualização, porém podemos salientar a arquitetura MultiSIG (STRAUCH, 1998) como uma proposta diferenciada na área ambiental. A MultiSIG propõe uma federação de dados com uma metodologia para integrar dados espaciais e uma arquitetura com esquema global para disponibilizar os dados. A MultiSIG resolve os conflitos de integração na elaboração do esquema global. A desvantagem desta proposta é quando ocorre a alteração dos esquemas locais ou a entrada de um novo esquema local, tornando-se necessário alterar o esquema global e rever as integrações.

4 A PROPOSTA X-ARC

Este capítulo apresenta a X-ARC quanto aos seus objetivos, suas características e funcionalidades pretendidas, seus componentes, a abordagem de implementação, os serviços oferecidos, os papéis dos integrantes, o protocolo de registro de integrantes e o modelo de dados adotado pela arquitetura.

4.1 Arquitetura Proposta

A X-ARC foi projetada para acessar repositórios de dados estruturados (armazenados em SGBD relacionais, SGBD orientados a objetos, etc), semi-estruturados (documentos XML, arquivos texto, planilhas eletrônicas, etc) e espaciais (armazenados em SIGs). Basta que para cada repositório de dados exista um tradutor que realize o mapeamento do modelo de dados do repositório para o modelo de dados comum do mediador.

A interoperabilidade dos dados só é possível devido aos metadados, pois representam a informação contida nos repositórios de forma coerente com os termos compreendidos pelo domínio dos usuários de dados ambientais e descrevem a estrutura e semântica existente nos dados.

A Figura 4.1 apresenta uma visão geral da arquitetura, com os gerentes e os participantes interligados pela Internet, juntamente com os repositórios de dados disponibilizados pelos mesmos.

Nesta Figura, é possível observar que um mesmo Participante pode publicar um ou mais repositórios de dados e que um Gerente também pode publicar dados diretamente. Também é possível notar que cada Participante tem um conjunto de metadados locais associado, que é capturado pela arquitetura durante a fase de

publicação do dado pelo usuário publicador e gerenciado pelo Participante, e opcionalmente um conjunto de metadados remotos. Os Gerentes também possuem um conjunto de metadados, contudo estes são globais e consolidam todos os metadados locais e encontram-se replicados entre os Gerentes da arquitetura.

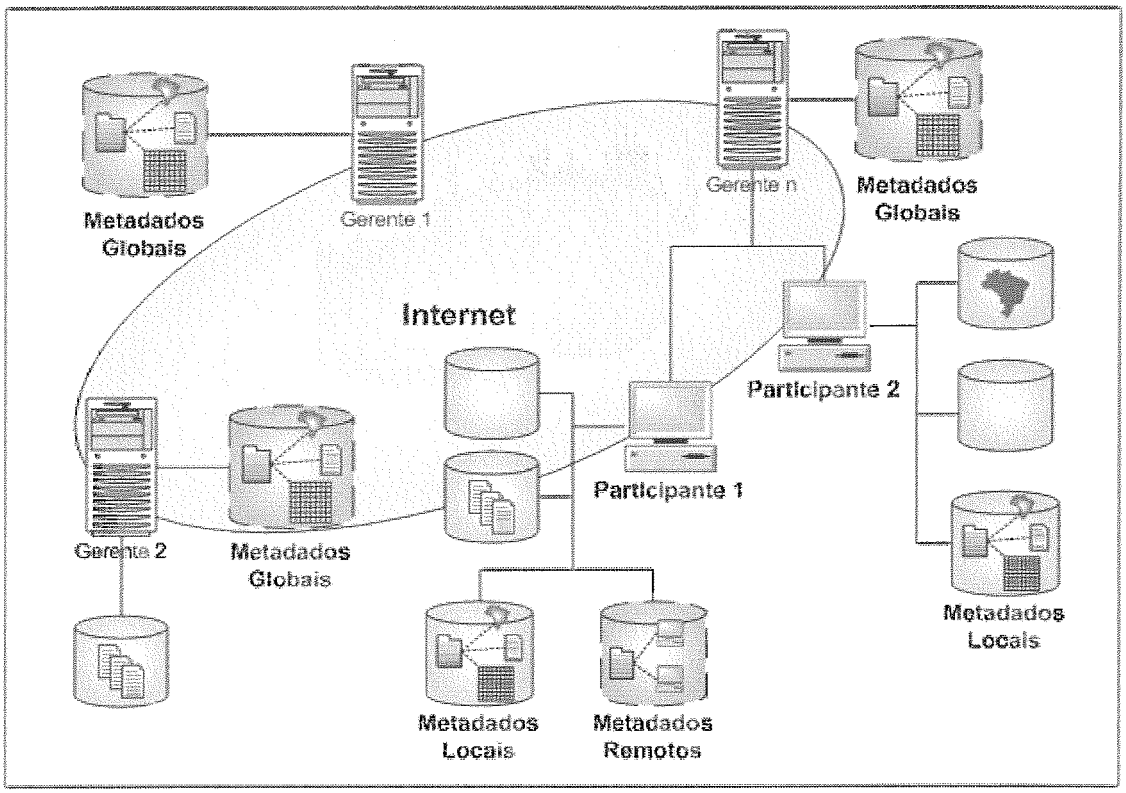


Figura 4.1 – Visão Geral da Arquitetura X-ARC

A arquitetura X-ARC encapsula dois sistemas para a execução de seus serviços e apoio à publicação dos dados: o Le Select (LESELECT, 2002), *midleware* do INRIA que manipula dados estruturados e semi-estruturados; e o ArcIMS (ARCIMS, 2001) que manipula dados georreferenciados.

Na Figura 4.2, temos uma visão genérica dos componentes que participam da arquitetura X-ARC. Nesta figura, é possível observar o mediador que acessa os repositórios de dados disponibilizados pelos sistemas de publicação de dados espaciais (ArcIMS) e não-espaciais (Le Select). O X-SELECT e o X-MAP são os

componentes tradutores da arquitetura que encapsulam o acesso aos sistemas Le Select e ArcIMS.

A arquitetura X-ARC visando aproveitar o esforço e pesquisa envolvidos na área de banco de dados, utiliza o Le Select para prover os serviços de acesso a dados não-espaciais. Da mesma forma, o ArcIMS é utilizado para prover o acesso aos dados espaciais disponíveis através de mapas digitais. Inicialmente, a arquitetura fornece acesso para dados armazenados nos formatos *shape*, texto, planilha eletrônica Excel e banco de dados relacionais.

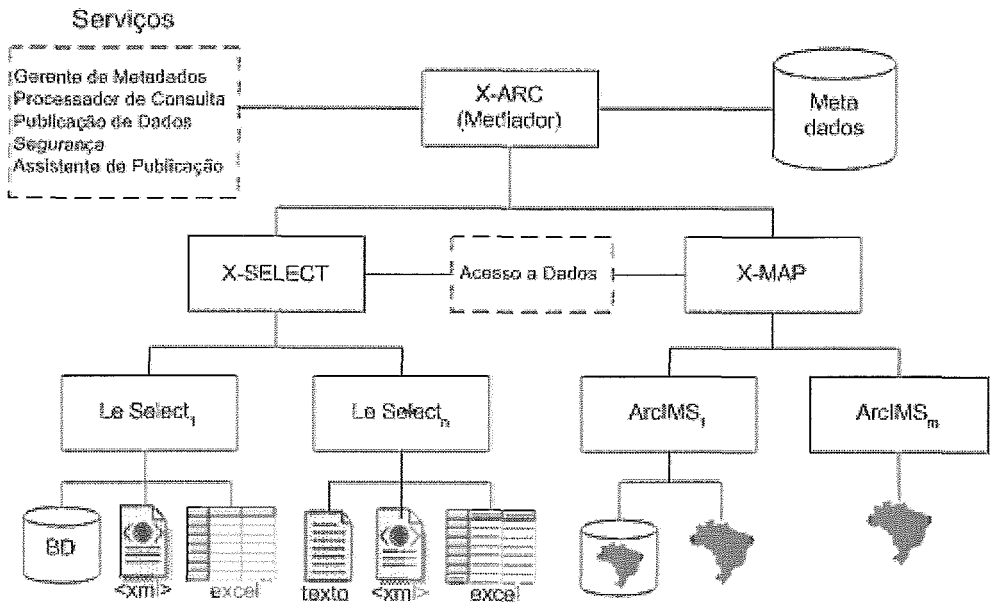


Figura 4.2 – Componentes da X-ARC

As características da arquitetura, os papéis assumidos por uma instância X-ARC, o modelo de dados do mediador, a gerência de metadados, os serviços disponíveis na arquitetura, o papel desempenhado pelo Le Select e pelo ArcIMS, assim como o processamento de consulta e a publicação dos dados disponibilizados pela X-ARC serão apresentados nas seções seguintes.

4.2 Características da Arquitetura

A fim de atender às necessidades dos usuários de Sistemas de Informação Geográfica (SIG) no acesso, localização e interoperabilidade de dados ambientais, indispensáveis para a realização de suas análises e para a tomada de decisões, foi desenvolvida a e**X**tensible **Arc**hitecture (X-ARC). A arquitetura X-ARC é baseada na arquitetura de mediadores (WIEDERHOLD, 1992) e é capaz de fornecer acesso a fontes de dados heterogêneas e distribuídas (PINTO, MEDEIROS, SOUZA et al., 2001, PINTO, STRAUCH, SOUZA et al., 2002a).

A X-ARC utiliza termos de domínio para apoiar a associação dos dados heterogêneos aos seus domínios de aplicação específicos e representar a informação existente nos repositórios de forma coerente com o vocabulário do domínio. Além de termos de domínio, a arquitetura utiliza a linguagem XML como um padrão de intercâmbio para disponibilizar os dados. Assim, os dados dos diversos repositórios envolvidos são acessados uniformemente pelos usuários e retornados como coleções as quais não apresentam heterogeneidade de formato pois são publicadas em XML, porém apresentam heterogeneidade semântica e esta deve ser resolvida pelo usuário quando decidir quais conjuntos de dados retornados na coleção utilizar.

Na X-ARC, cada fonte de dados é associada a um conjunto de metadados que fornece ao mediador as informações necessárias para o gerenciamento da arquitetura e o processamento das consultas, o que auxilia na identificação de fontes de dados correlatas e na localização de dados relacionados em um mesmo domínio de aplicação.

A X-ARC provê interoperabilidade entre os repositórios de dados, mantendo entretanto a autonomia de cada um dos repositórios participantes da arquitetura. Além disso, a disponibilidade e a capacidade de processamento de cada fonte, assim como a

heterogeneidade de seus dados (seja relativo à estrutura, formato, tipo ou metadados) são explorados pelos componentes da arquitetura a fim de garantir um acesso uniforme aos dados.

A arquitetura é flexível e escalável, pois novos repositórios de dados podem ser acrescentados à arquitetura sem muito esforço pelos usuários, através do uso do Assistente de Publicação de Dados, e sem redução no desempenho da arquitetura. Outra funcionalidade da arquitetura é a capacidade de auxiliar na localização de dados ambientais relacionados, utilizando a similaridade e equivalência de termos de domínio, associados aos repositórios de dados, na determinação de dados correlatos.

4.3 Mediador da X-ARC

Assim como em outras arquiteturas semelhantes que utilizam mediadores, a X-ARC é composta por um componente global (mediador) e por componentes locais (tradutores). O mediador acessa a informação compartilhada pelos componentes locais, gerenciando o acesso às fontes de dados locais e utilizando os tradutores que armazenam o mapeamento das fontes locais para o modelo de dados comum (MDC) do mediador.

A Figura 4.3 ilustra a comunicação e o fluxo de dados e metadados entre as camadas de aplicação, mediação, tradutores, sistemas intermediários e fontes de informação da arquitetura. A camada de aplicação representa um usuário ou aplicação que se conecta ao Gerente da arquitetura, envia consultas e aguarda o resultado. A camada de Mediação juntamente com a camada de Tradutores encontra-se no Gerente que recebe a consulta e a processa.

A camada de Sistemas Intermediários representa os possíveis servidores Le Select e ArcIMS que publicam os dados disponibilizados pela arquitetura e podem ser

acessados, respectivamente, através dos tradutores X-SELECT e X-MAP. A camada de fontes de informação representa os diversos repositórios distribuídos pela rede os quais são acessados através dos sistemas intermediários.

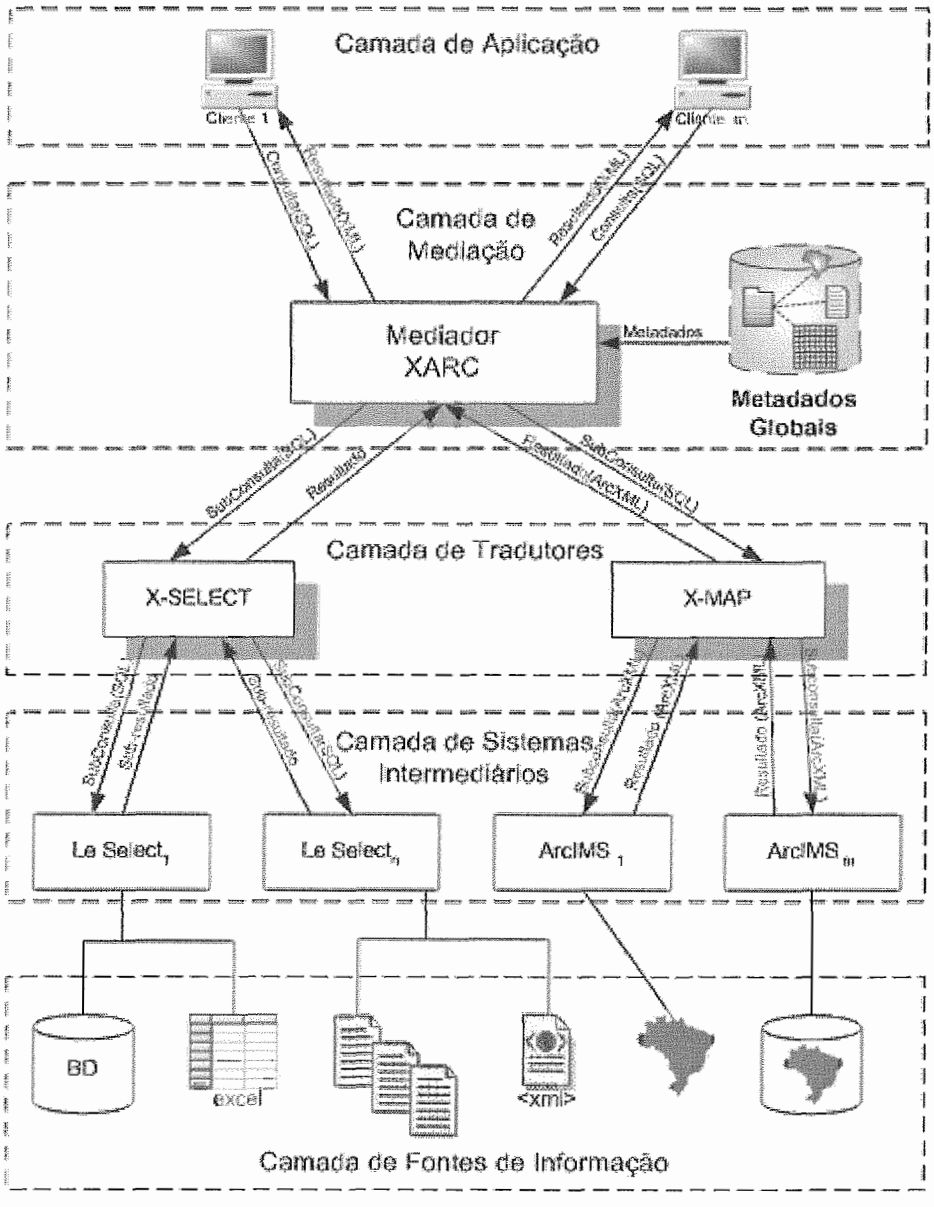


Figura 4.3 – Comunicação entre as camadas de Aplicação, Mediação, Tradução, Sistemas Intermediários e Fontes de Dados na Arquitetura X-ARC

A abordagem da visão do mediador é virtual pois os dados permanecem nos repositórios locais que são acessados através de consultas. As consultas enviadas ao mediador são transformadas em subconsultas (re-escritas nas linguagens de consulta

de cada repositório) e enviadas para os repositórios locais. Cada resultado é então traduzido, filtrado e unido em um único resultado que é retornado para o usuário ou aplicação que originou a consulta inicial.

O conceito de mediação empregado pela X-ARC para a publicação de dados diferencia-se do conceito tradicional de mediadores, pois a X-ARC não utiliza uma visão integrada dos repositórios no mediador, porém utiliza componentes globais (mediador) e componentes locais (tradutores) para publicar os dados disponibilizados pelos usuários.

A integração de dados ambientais é uma atividade mais complexa que a integração convencional de dados, pois os aspectos dos dados ambientais são mais complexos e semanticamente ricos. A integração destes dados necessita de constante intervenção humana e da presença de um especialista do domínio.

A interoperabilidade de dados ambientais distribuídos e heterogêneos é o objetivo da arquitetura, porém é necessário considerar a indisponibilidade temporária dos dados e/ou o fato que a decisão de como unir, transformar e utilizar os dados ambientais é responsabilidade dos usuários dos dados. Portanto, o problema de manutenção da visão do mediador fica evidente e proibitivo para um ambiente com repositórios de dados ambientais instáveis.

Desta forma, a arquitetura foi concebida para ao invés de integrar os dados, publicar o dados. A publicação dos dados é feita utilizando o conceito de mediação (o mediador gerencia o acesso aos dados disponibilizados pelos tradutores), entretanto a complexidade na manutenção da visão do mediador foi substituída pelo uso de termos de domínio que classificam e agrupam os dados em coleções e a adoção de metadados que descrevem os dados e auxiliam na sua localização.

4.4 Papéis na Arquitetura

Na arquitetura, uma instância X-ARC pode ter o papel de Gerente ou de Participante. Da mesma forma, existem diferentes denominações para os usuários que manipulam a arquitetura, sendo o Usuário Produtor e o Usuário Consumidor os mais importantes. Cada usuário ou instância da X-ARC desempenha funções diferentes na arquitetura. As características e funções de cada um são apresentados a seguir, enquanto que o diagrama em UML dos usuários da arquitetura encontra-se no Apêndice A.

4.4.1 Usuários

Os usuários que manipulam a arquitetura X-ARC são apresentados pela Figura 4.4 e divididos nas seguintes categorias:

- Usuário Administrador – responsável pela configuração da instância X-ARC no momento da instalação da instância em um computador. Ele configura na instância o perfil da instância (Participante ou Gerente) e fornece outras informações necessárias para a inicialização da instância na rede X-ARC;
- Usuário Consumidor – usuário que utiliza os serviços da arquitetura X-ARC para o acesso, localização e recuperação de dados disponíveis nos repositórios publicados pelos integrantes da arquitetura. Este usuário foi registrado na arquitetura pelo Usuário Administrador e tem esta denominação de acordo com a classificação de TOMASIC & SIMON (1997);
- Usuário Convidado – trata-se de um usuário semelhante ao Usuário Consumidor contudo não foi registrado pelo Usuário Administrador. Existe com o intuito de promover uma cooperação entre os usuários que

não participam da arquitetura, permitindo que este usuário realize consultas sobre os dados com nível de acesso Completo;

- Usuário Produtor/Publicador – usuário que disponibiliza dados para os integrantes da arquitetura. Este usuário interage nos integrantes, fornecendo informações (metadados) sobre os dados, tais como estrutura, descrição, qualidade, termo de domínio, etc. Este usuário tem esta denominação de acordo com a classificação de TOMASIC & SIMON (1997).

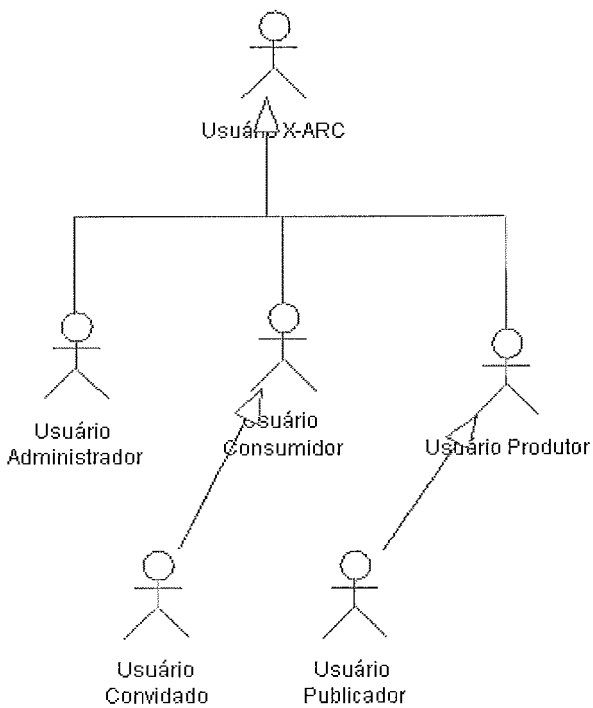


Figura 4.4 – Usuários da Arquitetura X-ARC

4.4.2 Participante

O Participante possui uma instância da X-ARC sendo executada localmente. Ele gerencia apenas seus metadados locais os quais descrevem as fontes de dados que disponibiliza. Opcionalmente pode possuir metadados remotos, porém não tem permissão para alterar as informações nestes metadados.

O Participante pode disponibilizar uma ou mais fontes de dados dentro da arquitetura. O tipo de fonte de dados que o Participante disponibiliza serve para classificá-lo, a saber:

- Espacial – disponibiliza dados espaciais (mapas) através de imagens; e
- Não-Espacial – disponibiliza dados armazenados em tabelas, textos, arquivos, bancos de dados, etc.;
- Híbrido – quando o participante disponibiliza dados espaciais e não-espaciais.

O Participante está sempre relacionado diretamente a um Gerente, a quem esta subordinado e tem como dever, informar seus metadados locais. Além do Gerente direto, um Participante pode possuir uma lista de gerentes alternativos, tratam-se dos gerentes existentes na rede X-ARC que foram informados da existência do Participante na rede. Para estabelecer como os metadados locais são informados ao Gerente e como a lista de gerentes alternativos é informada ao Participante é definido um protocolo apresentado na seção 4.5.

4.4.3 Gerente

O Gerente assim como o Participante possui uma instância da X-ARC sendo executada. Ele é responsável por gerenciar seus próprios metadados, os metadados locais dos participantes que estão subordinados a ele diretamente e os metadados dos outros integrantes da rede X-ARC (gerentes e participantes). A este conjunto de metadados controlados pelo Gerente denomina-se Metadados Globais. Embora possua os metadados referentes às fontes de dados disponibilizadas pela arquitetura, ele somente tem permissão para alterar seus metadados. Assim como o Participante, o

Gerente também pode disponibilizar uma fonte de dados, por exemplo o **Gerente 2** da Figura 4.1.

O Gerente pode estar relacionado diretamente a nenhum ou mais gerentes. Ele estará relacionado a nenhum quando for a única instância X-ARC na rede, ou seja, quando a rede estiver sendo iniciada. Embora exista um relacionamento direto entre os gerentes não existe uma hierarquia entre os mesmos.

Um Gerente possui uma lista de todos os gerentes existentes na rede. Além disso, possui uma lista dos participantes subordinados a ele e uma lista dos outros participantes da rede. Os metadados globais mantidos pelo Gerente são atualizados cada vez que uma fonte de dados é disponibilizada, ou um novo integrante se insere na rede X-ARC ou uma alteração nas fontes de dados disponibilizadas é feita. Os detalhes a respeito da troca de metadados globais entre os gerentes, a manutenção da lista de gerentes e a formação da rede X-ARC serão discutidos na seção 4.5.

4.5 Protocolo de Registro de Integrantes

Os integrantes da arquitetura X-ARC (participantes e gerentes) encontram-se espalhados pela Internet e se comunicam através dela para compor um conjunto que se denomina rede X-ARC.

A entrada de cada integrante na rede resulta na execução de atividades que devem ser desempenhadas pelo Usuário Administrador (pessoa responsável por instalar e configurar uma instância da X-ARC em um recurso computacional), pela própria instância X-ARC que está sendo inserida na rede e pelas outras instâncias que já se encontram ativas na mesma.

Tendo em vista que a X-ARC é uma arquitetura de mediação aberta, a qual permite a qualquer momento a entrada e saída de novos repositórios de dados, torna-

se necessário definir um protocolo de registro dos participantes na arquitetura. Este protocolo define um conjunto de procedimentos a serem realizados pelo usuário, pela nova instância que está entrando na rede e pelas instâncias X-ARC que já se encontram na rede.

Os procedimentos podem ser divididos entre os desempenhados pelos usuários da arquitetura e os desempenhados pelas instâncias X-ARC que compõem a rede. Os diagramas de Caso de Uso em UML encontram-se nos Apêndice A.

Os procedimentos de responsabilidade do Usuário Administrador constituem-se os seguintes:

- Instalação da instância X-ARC em um recurso computacional com acesso à rede Internet; e
- Configuração da instância quanto ao perfil de execução (gerente ou participante) e indicação, se possível, do endereço IP de outra instância X-ARC em execução.

Os procedimentos desempenhados pelas instâncias X-ARC durante o registro de uma nova instância na rede podem ser divididos entre os realizados pela nova instância e aqueles realizados pelas instâncias já existentes.

A nova instância deve a partir do número IP fornecido pelo Usuário Administrador iniciar uma comunicação com a instância já existente, informando a entrada de uma nova instância na rede. Neste momento, a nova instância deve informar seu papel (Gerente ou Participante) na rede X-ARC. Só então outras informações são trocadas entre as instâncias, tais como: lista de gerentes, metadados globais ou locais, etc.

Caso a nova instância seja um gerente, ela receberá uma lista dos gerentes existentes na rede, juntamente com os metadados globais. Se for um participante,

então uma lista com gerentes alternativos será informado. Esta lista não é atualizada a cada novo gerente que se insere na rede, mas serve como um recurso para a instância caso seu gerente saia da rede.

Durante a entrada de uma nova instância na rede, as instâncias já existentes na rede X-ARC devem verificar o papel da nova instância na rede, para então determinar os próximos passos a serem executados.

Se a nova instância for gerente, os metadados globais da instância já existente devem ser informados a ela e os gerentes da rede devem ser informados da entrada de um novo gerente. Caso seja participante, uma lista de gerentes alternativos deve ser informada. Se a nova instância indicou a instância durante seu processo de registro, então ela deve ser incluída em sua lista de participantes subordinados diretos. Caso contrário, deve apenas atualizar sua lista de participantes na rede.

4.5.1 Exemplo de Registro de Integrantes

A seguir, é apresentado um exemplo que ilustra a formação da rede X-ARC através do registro de seus integrantes. Algumas observações sobre os metadados manipulados pela instância são fornecidos para o esclarecimento de dúvidas que tenham surgido durante a seção anterior. O exemplo é apresentado em cenários ao longo de diferentes instantes de tempo e com variadas configurações da rede X-ARC. Este exemplo procura ilustrar as diferentes configurações que a rede pode assumir e as iterações entre as instâncias X-ARC durante o registro dos integrantes.

A Figura 4.5 apresenta o Cenário 1 (Instante T=0), no qual o integrante P1 inicia a rede X-ARC, só existe ele na rede, portanto ele é Gerente e Participante ao mesmo tempo.

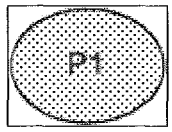


Figura 4.5 – Registro de Integrantes (Cenário 1)

A Figura 4.6 apresenta o Cenário 2 (Instante T=1) onde a partir da rede já formada pelo integrante P1 do cenário anterior, o integrante P2 se registra em P1 como Participante. Neste instante, é formada a rede com 1(um) Gerente e 1(um) Participante.

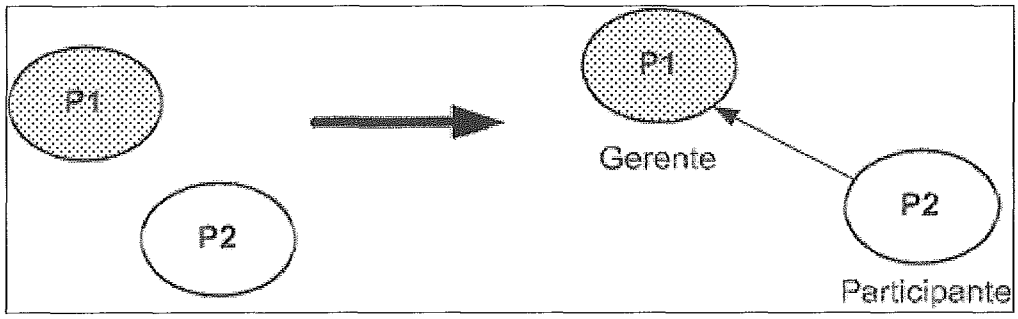


Figura 4.6 – Registro de Integrantes (Cenário 2)

A Figura 4.7 apresenta o Cenário 3 (Instante T=1'), onde a partir da rede já formada pelo integrante P1 da Figura 4.5 (Cenário 1), o integrante P2 se registra em P1 como Gerente. Neste instante, é formada a rede com dois gerentes.

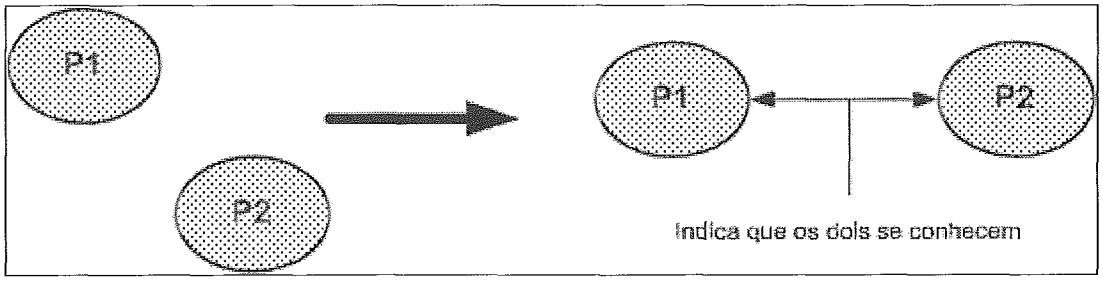


Figura 4.7 – Registro de Integrantes (Cenário 3)

A Figura 4.8 apresenta o Cenário 4 (Instante T=2), onde a partir da rede do cenário anterior, o integrante P3 se registra em P2 como Participante. Neste caso, o

integrante P1 é informado da existência de P3, através da atualização dos metadados enviada por P2. O integrante P3 é informado da existência de P1, como gerente alternativo, durante seu registro com P2.

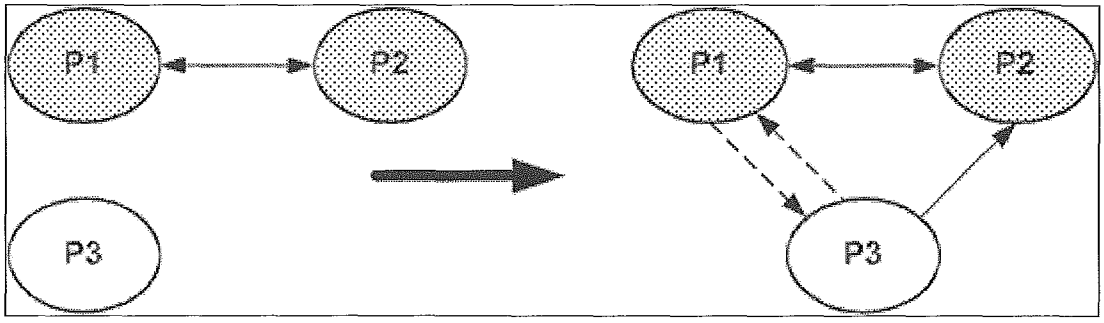


Figura 4.8 – Registro de Integrantes (Cenário 4)

A Figura 4.9 apresenta o Cenário 5 (Instante T=2'), onde a partir da rede da Figura 4.7 (Cenário 3), o integrante P3 se registra em P2 como Gerente. Neste caso, P1 é informado da existência de P3, através da atualização dos metadados enviada por P2. P3 é informado da existência de P1 (gerente), durante seu registro com P2. Os metadados globais existentes em P2 são enviados para P3.

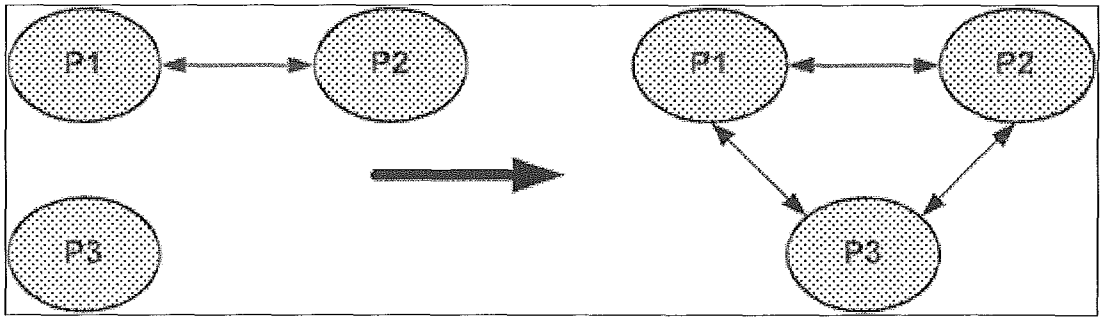


Figura 4.9 – Registro de Integrantes (Cenário 5)

A Figura 4.10 apresenta o Cenário 6 (Instante T=3), onde a partir da rede do cenário anterior, o integrante P4 se registra em P3 como Participante. Neste caso, P1 e P2 são informados da existência de P4, através da atualização dos metadados enviada por P3. P4 é informado da existência de P1 e P2 (gerentes alternativos), durante seu registro com P3.

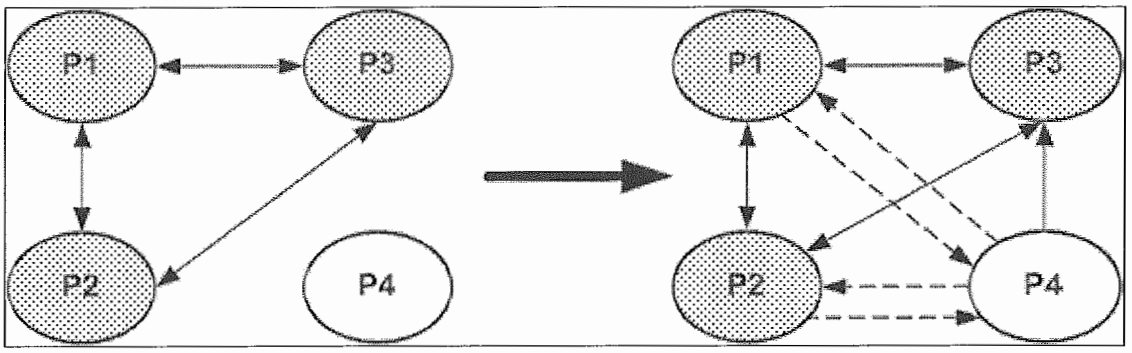


Figura 4.10 – Registro de Integrantes (Cenário 6)

A Figura 4.11 apresenta o Cenário 7 (Instante T=3'), onde a partir da rede da Figura 4.7, o integrante P4 se registra em P2 como Participante. Neste caso, P1 é informado da existência de P4, através da atualização dos metadados enviada por P2. P4 é informado da existência de P1 (gerente alternativo), durante seu registro com P2. Os integrantes P3 e P4 não sabem da existência um do outro.

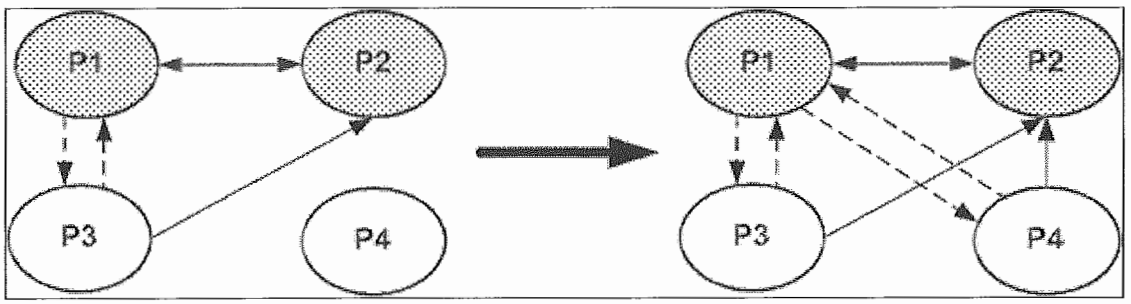


Figura 4.11 – Registro de Integrantes (Cenário 7)

4.6 Modelo de Dados

A X-ARC como uma arquitetura de publicação de dados que baseia-se no conceito da arquitetura de mediadores requer a definição de um modelo de dados comum (MDC) ou canônico para o mediador da arquitetura.

O MDC permite que os usuários realizem consultas sobre os vários esquemas participantes da mediação utilizando uma linguagem de consulta padrão, independente das linguagens de consulta disponibilizada pelos repositórios. Cabe ressaltar que o tradutor de cada repositório tem a responsabilidade de traduzir a consulta na linguagem do mediador para uma consulta na linguagem do repositório de dados.

Segundo (BUSSE, KUTSCHE, LESER et al., 1999), o modelo de dados das arquiteturas de integração restringe os tipos de componentes (repositórios de dados) que podem ser integrados, pois durante a tradução de um modelo (relacional, semi-estruturado, orientado a objetos, etc.) para outro pode haver perda de representação semântica. Além disso, a linguagem de consulta aos dados é de fundamental importância, pois uma linguagem complexa cria barreiras de utilização, podendo inviabilizar o uso da arquitetura pelos usuários.

Avaliando estes critérios e considerando que na área ambiental a maioria dos usuários não é especialista em banco de dados, a X-ARC utiliza o modelo relacional como seu modelo de dados canônico e um subconjunto da linguagem SQL como sua linguagem de consulta padrão. O modelo relacional foi adotado devido a facilidade de conversão dos dados ambientais representados em outros modelos para o relacional e ao fato que o modelo relacional ser comumente utilizado pela indústria, além da representação tabular ser comum aos usuários da área ambiental que estão acostumados com a manipulação de dados neste formato.

O modelo de mediação da X-ARC não visa gerar um esquema de integração único. As técnicas de mediação são utilizadas para especificar um mapeamento entre os tipos do mediador e os tipos dos repositórios de dados locais e prover um acesso lógico unificado aos dados existentes nos repositórios participantes. Desta forma, o resultado de uma consulta a dados disponíveis em vários repositórios é retornado como uma coleção de dados, onde cada dado possui sua estrutura e semântica própria, mas em um único formato, um único documento XML bem formado.

A X-ARC disponibiliza os dados existentes nos repositórios de dados através de uma abstração dos repositórios baseado no conceito de *site* publicado, onde cada repositório é representado por uma *Unified Resource Locator* (URL) que descreve o computador e o repositório, ou seja, as fontes de dados tornam-se acessíveis para qualquer aplicação cliente que possa se conectar à X-ARC e informar uma URL válida.

Através da abordagem de *site* publicado utilizado pela X-ARC, encapsula-se os dados publicados pelos sistemas intermediários ArcIMS e Le Select, onde o ArcIMS disponibiliza os dados espaciais na Web como figuras que são geradas pelo sistema de acordo com a demanda específica (ARCIMS, 2001). Da mesma forma, o Le Select disponibiliza seus dados e serviços através do conceito de *site* publicado (XHUMARI, AMZAL & SIMON, 1999).

Essa abordagem dificulta a tarefa de localização de informações dos repositórios caso não exista um catálogo de repositórios publicados, entretanto a X-ARC propõe um Serviço de Gerente de Metadados que tem como uma de suas funções, fornecer as informações sobre os repositórios publicados.

4.7 Metadados

A gerência dos metadados é um fator preponderante para o desempenho dos serviços da arquitetura, que auxiliam no acesso, gerenciamento e compartilhamento de grandes conjuntos de dados estruturados e não-estruturados. Através de uma correta gerência se atinge a interoperabilidade entre as diversas fontes de dados manipuladas pela arquitetura.

Devido à heterogeneidade e distribuição das fontes de dados publicadas pela arquitetura, torna-se extremamente necessário a descrição de informações a respeito dos repositórios de dados e dos próprios dados, tanto em relação ao tipo de armazenamento, como também a estrutura, forma de acesso, etc.

A definição de um conjunto de metadados adequado para a arquitetura X-ARC é o primeiro passo para garantir uma implementação eficiente no acesso, transferência, compartilhamento e processamento dos dados. Além disso, na arquitetura de mediadores, os metadados auxiliam o mediador na determinação da execução das tarefas da arquitetura.

Os metadados da X-ARC são controlados pelo Gerente de Metadados (GM) que fornece, coleta e gerencia as informações sobre os repositórios. O GM apoia os outros serviços da arquitetura, como por exemplo a decomposição de consultas, o empacotamento de resultados, etc. É também através do GM que o usuário tem a facilidade de consultar as informações existentes sobre os diversos repositórios de dados publicados pela arquitetura, sem conhecer ou compreender a estrutura dos dados nos repositórios locais.

Para cada repositório que é incorporado à arquitetura, o GM captura um conjunto de informações que descrevem o tipo do dado (espacial ou não espacial), a localização (computador que armazena), forma de armazenamento, estrutura do dado,

qualidade do dado, etc. Todas essas informações são fornecidas pelo usuário durante o processo de publicação dos dados. O critério de qualidade utilizado pela arquitetura é subjetivo, depende do usuário, podendo no entanto ser aprimorado utilizando técnicas específicas.

Como foi citado no capítulo 2, existem várias arquiteturas e padrões de metadados, alguns visam domínios específicos de aplicação, enquanto outros procuram ser genéricos. Considerando a heterogeneidade dos repositórios envolvidos, várias arquiteturas ou padrões poderiam ser utilizadas para descrever os dados disponibilizados, todavia não é foco principal deste trabalho desenvolver de forma completa uma gerência de metadados, embora seja possível estender a arquitetura para tal objetivo. Portanto, o modelo de metadados apresentado a seguir terá como principal propósito, o de representar as necessidades da arquitetura para a execução de suas tarefas.

Para a X-ARC foi adotado um conjunto de metadados próprio para as atividades da arquitetura. Entretanto, podem ser estendidos para seguir um padrão, arquitetura ou formato de metadados como FGDC, Dublin Core ou SAIF.

A Figura 4.12 apresenta o diagrama de classes, na notação UML (BOOCH, RUMBAUGH & JACOBSON, 2000), dos dados armazenados pelos metadados da X-ARC e manipulados pelo Gerente de Metadados (GM).

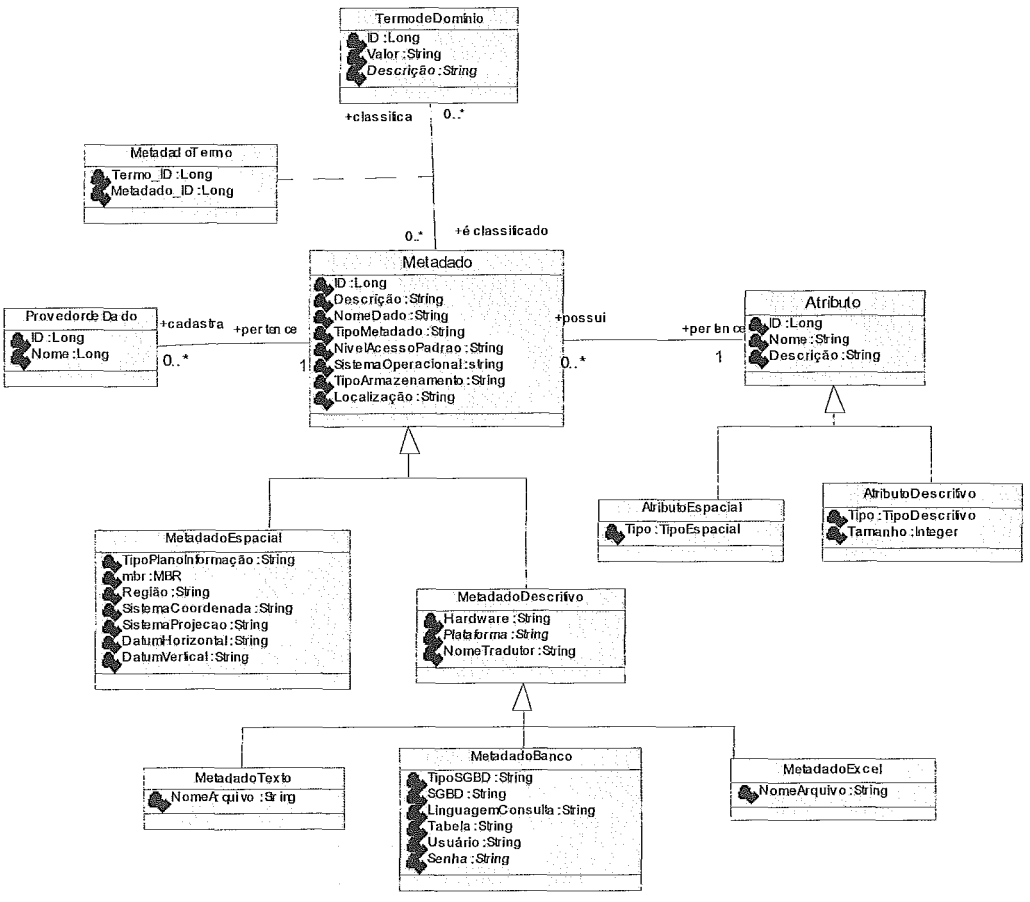


Figura 4.12 – Diagrama de Classes dos Metadados

4.7.1 Tipos de Metadados

Os metadados armazenados e gerenciados pela arquitetura classificam os dados disponibilizados pelos repositórios em dois tipos: espaciais e não-espaciais. Além disso, os metadados utilizados pela arquitetura são classificados em três grupos, a saber:

- Local – são os metadados mantidos pelos participantes e pelos gerentes da arquitetura. Para cada dado disponibilizado pelo participante/gerente existe um metadado associado que o descreve e informa o nível de acesso autorizado pelo usuário que disponibilizou o dado. Somente o

participante/gerente que publica o dado tem permissão para alterar os metadados relativos ao mesmo;

- Remoto – são os metadados que descrevem um dado disponibilizado por outro participante/gerente da arquitetura. É semelhante ao metadado local, porém o participante que o possui não tem autonomia para alterá-lo. Este metadado é um recurso que a X-ARC oferece aos participantes para estender seu conjunto de metadados e diminuir sua dependência do serviço de metadados dos gerentes;
- Global – são os metadados mantidos pelos gerentes da arquitetura. Estes metadados descrevem os dados disponibilizados por todos os participantes que compõem a arquitetura, ou seja, é o conjunto dos metadados locais de todos os participantes. Os gerentes não têm permissão para alterar a informação contida nos metadados globais, preservando a autonomia de projeto dos participantes durante o processo de publicação dos dados.

4.7.2 Protocolo de Intercâmbio de Metadados

Os metadados constituem informação tão relevante quanto o próprio dado que eles descrevem e por isso a arquitetura deve ter um protocolo bem definido sobre o intercâmbio dos metadados entre os integrantes (participantes e gerentes) da arquitetura.

O intercâmbio de metadados na arquitetura ocorre principalmente em dois casos: durante o registro de um novo integrante (gerente ou participante) na arquitetura e durante a execução de consultas, seja para consultar possíveis repositórios ou para recuperar os dados existentes nestes.

Durante o processo de registro de integrantes na arquitetura, os participantes trocam metadados com seus gerentes e estes trocam metadados entre si para atualizar seus metadados globais sobre os novos dados disponibilizados pelos participantes.

Na execução das consultas, as informações solicitadas pelo cliente que enviou a consulta são retornadas acrescidas de metadados adicionais, como estrutura, descrição, qualidade, etc.

Para estes casos a arquitetura deve fornecer segurança dos dados e metadados a partir de dois itens muito importantes:

- Nível de Acesso – garantindo que os dados e metadados publicados estejam de acordo com o nível de acesso atribuído por seus usuários. A arquitetura adota o mesmo nível de acesso atribuído aos dados pelos usuários para os seus respectivos metadados; e
- Autenticação – garantindo que os metadados trocados entre os participantes e gerentes, e estes entre si, estejam protegidos durante a comunicação. Esta proteção poderia ser obtida através de técnicas de criptografia de informação como criptografia forte, criptografia convencional, chaves públicas e chaves privadas. Além disso, as informações transmitidas poderiam ser autenticadas através de assinaturas digitais, as quais permitem ao receptor verificar a autenticidade da origem da informação e de que permanece intacta (RSA LABORATORIES, 2000 , RIVEST, SHAMIR & ADLEMAN, 1978, RIVEST, 1992).

4.8 Serviços da arquitetura

Em cada um dos integrantes da arquitetura (gerente e participante) encontra-se um ou mais serviços que são responsáveis pela localização, integração ou gerenciamento dos repositórios manipulados pela arquitetura, conforme ilustrado pela Figura 4.2. A Figura 4.13 apresenta os serviços atualmente disponíveis na arquitetura e suas interações. Estes serviços compõem a camada de integração de dados nas arquiteturas do projeto SPeCS e AGROMET, apresentados no capítulo 1. A seguir, a explicação sobre cada um dos serviços e suas funções será apresentado.

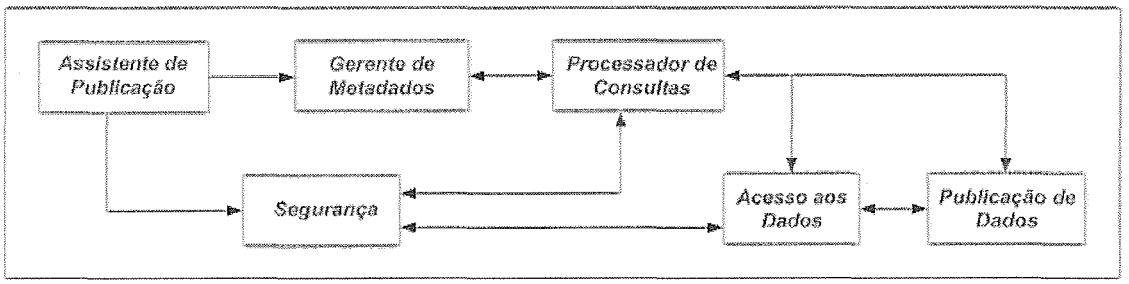


Figura 4.13 – Serviços da Arquitetura X-ARC

4.8.1 Assistente de Publicação

A participação dos repositórios de dados na arquitetura, deve partir da iniciativa dos usuários em partilhar dados que detém. Uma vez que a tecnologia utilizada para a publicação dos dados é a mediação, torna-se necessário para cada novo repositório incluído na arquitetura, a criação de um tradutor específico para o mesmo. Assim sendo, é necessário que a arquitetura forneça ao usuário uma funcionalidade para auxiliá-lo na criação do tradutor para o dado a ser incluído. Esta é a principal função do Assistente de Publicação (AP).

O AP é responsável por auxiliar os usuários, que por ventura desejam disponibilizar seus dados na arquitetura, na construção de um tradutor específico para seus dados. Para dados não-espaciais o AP auxilia na montagem do arquivo de definição de tradutor (*wrapper definition file*), que são utilizados pelo Le Select para publicar os repositórios de dados. Também é de responsabilidade deste serviço, o registro da definição do tradutor no servidor Le Select que o participante da arquitetura possui.

Além disso, durante o processo de publicação do dado, o AP repassa ao Gerente de Metadados os metadados sobre o novo dado disponibilizado. De acordo com o tipo de dado a ser publicado, espacial ou descritivo, e a forma como o dado está armazenado, tradutores específicos serão gerados. A partir deste serviço a publicação dos dados pelos usuários torna-se mais simples.

4.8.2 Segurança

O serviço de segurança é responsável por garantir a segurança de acesso aos dados e metadados disponibilizados pelos nós participantes da arquitetura, ou seja, o nível de acesso atribuído pelo usuário no momento da publicação do dado é aplicado também ao metadado que o descreve. Assim sendo, de acordo com os três níveis de acesso abaixo, os dados e metadados são disponibilizados pela arquitetura sob os seguintes critérios:

- Completo – os dados e os metadados estão disponíveis aos usuários para visualização e para extração;
- Somente Metadado – embora os dados e metadados componham a arquitetura, os dados não estão disponíveis para visualização. A existência

dos dados na arquitetura podem ser indicados a usuários através da visualização simplificada dos metadados que descrevem o dado. Portanto os dados apenas podem ser indicados como existentes, mas não podem ser visualizados ou extraídos;

- Negado – os dados e metadados fazem parte da arquitetura, porém não estão disponíveis para acesso.

O nível de acesso aos dados é atribuído pelo usuário que disponibilizou o dado, durante a execução do Assistente de Publicação. O serviço de segurança fornece ao serviço de Acesso aos Dados as informações sobre a permissão de acesso sobre cada dado a ser acessado.

A X-ARC não permite a realização de alterações nas bases de objetos componentes da arquitetura, uma vez que só tem permissão de leitura nas bases de dados locais, garantindo, assim, segurança local dos repositórios.

A confiabilidade e segurança dos metadados e dos dados disponibilizados pela arquitetura é alcançada através da aplicação do protocolo de intercâmbio de metadados definido na seção 4.7.2.

4.8.3 Gerente de Metadados

O Gerente de Metadados (GM) é responsável por gerenciar os metadados sobre todas as fontes de dados manipuladas pela arquitetura. As principais informações gerenciadas pelo GM sobre as fontes são: localização, tipo do dado (espacial, descritivo), forma de armazenamento, estrutura do dado e termo de domínio relacionado. A partir das informações mantidas pelo GM é possível consultar os dados existentes nos repositórios dos participantes segundo vários critérios de consulta.

O GM possui um dicionário de termos de domínio, que pode ser atualizado a qualquer momento, e é utilizado pelo Assistente de Publicação (AP) da arquitetura durante o processo de registro dos repositórios participantes. A partir do termo de domínio atribuído ao repositório, as fontes de dados que compõem a arquitetura são categorizadas segundo o domínio da aplicação e através desta classificação a X-ARC é capaz de auxiliar os usuários na localização dos dados.

4.8.4 Processador de Consulta

É responsável pelo processamento das consultas enviadas à arquitetura. Neste serviço a consulta é desmembrada e re-escrita em subconsultas específicas para cada repositório de acordo com os metadados globais mantidos pela arquitetura e fornecidos pelo GM. Cada subconsulta é enviada para o serviço de Acesso ao Dado que utilizando o tradutor adequado para o tipo de dado sendo consultado o acessa e o resultado de cada subconsulta é unificado em um único resultado pelo serviço de Publicação de Dados. Então este resultado unificado é enviado para o cliente que originou a consulta inicial.

Na X-ARC, os dados estruturados, semi-estruturados ou não-estruturados disponíveis nas fontes são acessados através de uma linguagem de consulta padrão (subconjunto da SQL) que permite a abstração do usuário sobre as diferenças de capacidade de execução de cada um dos repositórios envolvidos.

Maiores detalhes sobre o processamento de consultas na arquitetura, o subconjunto da SQL utilizado pela arquitetura e processo de como o resultado da consulta é retornado ao usuário ou aplicação que enviou a consulta serão discutidos nas seções 4.9 e 4.10.

4.8.5 Acesso aos Dados

É responsável pelo acesso direto aos dados publicados pelos participantes da arquitetura. A tarefa deste serviço é de acessar os dados disponibilizados pelos participantes e de repassá-los para o serviço de publicação de dados para que sejam publicados. Este serviço encapsula os serviços de mediação do Le Select e do ArcIMS e utiliza para se conectar com os mesmos os componentes X-Select e X-Map da arquitetura que serão descritos a seguir.

4.8.5.1 X-Select – eXtensible Le Select

Este componente da X-ARC é o responsável pela mediação de dados não-espaciais, todas as consultas direcionadas a repositórios não-espaciais são direcionados e tratados por este componente. Trata-se de um tradutor para repositórios de dados publicados pelo Le Select.

Além de ser responsável por traduzir as consultas enviadas pela arquitetura X-ARC para a linguagem de consultas do Le Select, também é responsável pela tradução do resultado retornado pelo Le Select para o formato manipulado pela X-ARC.

O tradutor X-SELECT se comunica com o Le Select através de um *driver* JDBC específico, já existente, fornecido juntamente com o Le Select. Este *driver* JDBC e as classes Java para manipulação dos resultados retornados pelo servidor são encontrados em um arquivo JAR, que é incluído no projeto da arquitetura para permitir a conexão com o servidor Le Select.

Na arquitetura X-ARC, vários servidores Le Select podem estar distribuídos entre computadores que formam uma rede de interesse em comum. Para cada

repositório de dados não-espacial que participa da publicação, existe pelo menos um servidor Le Select que o publique, entretanto um servidor pode ser responsável pela publicação de vários repositórios. Além disso, um mesmo repositório de dados pode ser disponibilizado por mais de um servidor Le Select, entretanto é necessário que um arquivo de definição do tradutor seja construído para o repositório e registrado em cada um dos servidores.

Como reflexo da utilização do Le Select, a linguagem de consulta disponibilizada por este componente é um subconjunto da linguagem SQL, que também é utilizada como linguagem de consulta da X-ARC.

A arquitetura X-ARC visa aproveitar a solução Le Select já existente, acrescentando sobre este sistema intermediário, uma camada para a execução das tarefas necessárias para a interoperabilidade de dados.

Embora já existam tradutores preestabelecidos para alguns tipos de repositórios de dados, a abordagem de implementação do Le Select permite que novos tradutores sejam construídos para atender às necessidades dos usuários. Este foi um dos fatores que contribuíram para a escolha deste sistema para compor a arquitetura X-ARC, pois na área ambiental é comum o uso de formatos, padrões e tipos de armazenamento proprietários.

Além disso, à medida que novos tradutores são desenvolvidos para diferentes repositórios manipulados por usuários de sistemas ambientais, eles podem ser distribuídos entre os usuários que utilizam a X-ARC, possibilitando um reaproveitamento de esforço.

Atualmente, todos estes aspectos são de fundamental importância para garantir que as soluções mantenham-se atualizadas, competitivas e configuráveis às necessidades específicas de cada usuário.

4.8.5.2 X-Map – eXtensible Map

Fazendo uma analogia com o componente X-SELECT, o X-MAP é o responsável pela mediação de dados espaciais. Todas as consultas direcionadas a repositórios espaciais são direcionados e tratados por este componente.

O X-MAP é o tradutor da X-ARC para repositórios de dados publicados pelo ArcIMS. Ele é composto por um cliente da arquitetura ArcIMS (ARCIMS, 2001), que foi concebida para servir dados e serviços de Sistemas de Informação Geográfica através da Internet (HARDER, 1998) . Por meio deste componente, é feito o acesso aos dados espaciais disponibilizados por um ou mais servidores ArcIMS. A Figura 4.14 ilustra o papel do X-MAP no acesso aos dados espaciais e como é realizada a interação do componente com os servidores ArcIMS.

Figura 4.14 – Papel do X-MAP no acesso aos dados espaciais

A arquitetura ArcIMS utiliza a ArcXML, uma versão da XML para o ArcIMS, para processar consultas e retornar resultados de consultas. Desta forma, o X-MAP é responsável pela tradução da consulta SQL para a linguagem ArcXML e pela conversão do resultado recebido em ArcXML para o formato da X-ARC. A Figura 4.15 exibe o trecho de uma consulta básica em ArcXML, no qual solicita uma imagem, que representa um conjunto de dados espaciais.

```
<ARCXML version="1.1">          // identificação da versão da ArcXML
<REQUEST>                       // indica que se trata de um pedido
<GET_IMAGE>                     // solicitação de uma imagem
<PROPERTIES>                    // informações adicionais sobre a imagem solicitada
<ENVELOPE minx="-180" miny="-90" maxx="180" maxy="90" />
</PROPERTIES>
</GET_IMAGE>
</REQUEST>
```

Figura 4.15 – Exemplo de Consulta em ArcXML

Neste exemplo, uma imagem que representa o mapa é solicitada. As coordenadas de interesse da região estão sendo fornecidas pelos valores de minx, miny, maxx e maxy (mínimo retângulo envolvente). Várias informações podem ser recuperadas utilizando marcadores específicos.

Na Figura 4.16 é apresentado um resultado em ArcXML retornado pelo servidor ArcIMS de uma consulta enviada ao mesmo.

```
<ARCXML version="1.1">          // identificação da versão da ArcXML
<RESPONSE>              // indica que se trata de uma resposta
<IMAGE>
// marcador que indica que foi retornado uma imagem para uma consulta
<PROPERTIES>            // informações adicionais sobre a imagem retornada
// informação sobre as coordenadas englobadas pela imagem retornada
<ENVELOPE minx="-180" miny="-90" maxx="180" maxy="90" />
<OUTPUT file="g:\arcims\output\world_MYCOMPUTER2102209.jpg"
url="http://mycomputer.domain.com/output/ world_MYCOMPUTER2102209.jpg" />
// nome e localização da imagem gerada pelo servidor
</PROPERTIES>
</IMAGE>
</RESPONSE>
</ARCXML>
```

Figura 4.16 – Exemplo de Resultado em ArcXML

É tarefa do X-MAP, realizar a conversão da consulta em formato SQL manipulado pela arquitetura X-ARC para a ArcXML e depois do formato ArcXML para um formato de resultado manipulado pela X-ARC.

Através do X-MAP, vários servidores ArcIMS podem ser acessados, basta que os dados que estes servidores disponibilizam estejam registrados pelos usuários no banco de metadados da X-ARC.

No âmbito da tecnologia de Geoprocessamento, não existe uma linguagem de consulta padrão como a SQL (*Structured Query Language*) para bancos de dados relacionais. Uma linguagem de consulta espacial requer a combinação de diversos fatores relacionados à geometria computacional, álgebra de objetos e técnicas de análise espacial (STRAUCH, 1998). Segundo RAMIREZ (2001) a questão referente à linguagem de consulta para sistemas espaciais permanece em discussão. Em seu trabalho, ele cita algumas pesquisas e propostas que têm sido feitas.

Assim, o componente X-MAP faz a tradução da linguagem padrão da arquitetura para a linguagem de consulta espacial manipulada pelo ArcIMS.

4.8.6 Publicação de Dados

O serviço de Publicação de Dados (PD) é responsável pela publicação dos dados existentes em cada participante da arquitetura em um formato que possibilite o intercâmbio de dados. Atualmente, o PD utiliza a linguagem XML como a linguagem de intercâmbio de dados padrão, uma vez que sua estrutura é flexível e capaz de se adaptar aos diversos formatos dos dados.

O PD publica os dados disponibilizados pela arquitetura de duas formas, de acordo com a natureza do dado. Para dados descritivos, o PD cria um documento

XML "bem formado" e insere neste documento o conteúdo (tuplas) retornado pelo serviço de Acesso aos Dados (maiores detalhes serão apresentados adiante). No caso dos dados espaciais, estes serão publicados como uma imagem. Assim sendo, um documento XML "bem formado" também é criado, no entanto ao invés de inserir as tuplas dos dados espaciais no documento, uma referência à imagem que os representa é incluída no documento (maiores detalhes serão enfocados na seção Publicação de Dados).

4.9 Processamento de Consultas

O processamento de consultas é considerado tema central na área de banco de dados (RAMIREZ, 2001, ELMASRI & NAVATHE, 2000) e representa na arquitetura umas das atividades mais importantes.

4.9.1 Linguagem de Consulta

A linguagem de consulta da X-ARC é um subconjunto da linguagem padrão SQL, onde as cláusulas de agrupamento (GROUP BY), ordenação (ORDER BY) e as funções de agregação como SUM, AVG, MIN e MAX entre outras foram retiradas e a cláusula espacial BY REGION foi acrescentada. O subconjunto SQL manipulado pela arquitetura permite as seguintes cláusulas:

- **SELECT** – seguida pelo caracter "*" ou por uma seqüência de nomes de colunas intercaladas pelo caracter ",". O nome das colunas só pode ser utilizado quando a consulta indicar na cláusula "FROM" um único repositório;
- **FROM** – seguida por uma URL que identifica um repositório disponibilizado por algum integrante da X-ARC ou seguida por um termo de domínio antecedido pelo caracter "<" e sucedido pelo caracter ">", por exemplo "<Precipitação>";
- **WHERE** – semelhante ao SQL padrão, onde indica a condição de seleção dos registros a serem recuperados. Esta cláusula pode ser utilizada somente quando a consulta indicar na cláusula "FROM" um único repositório; e

- **BY REGION** – indica que para repositórios espaciais deve ser retornado apenas os dados espaciais que se encontram localizados na região determinada pelo retângulo mínimo envolvente que segue a cláusula.

Alguns exemplos de consultas válidas enviadas à arquitetura são:

1. **SELECT** * **FROM** <TermodeDomínio>;
2. **SELECT** Coluna1, Coluna2 **FROM** //192.168.0.1/Estacao/Estacao1;
3. **SELECT** Coluna1, Coluna2 **FROM** //192.168.0.1/Estacao/Estacao1
   WHERE Coluna1 = 10;
4. **SELECT** * **FROM** //192.168.0.2/EstadoRJ
   BY REGION xMin,yMin,xMax,yMax;

4.9.2 Execução de Consulta na X-ARC

No caso de repositórios de dados não publicados pela arquitetura X-ARC, os usuários são obrigados a realizar as seguintes tarefas antes que possam utilizar um dado: encontrar os dados relevantes entre os diversos repositórios existentes; compreender a estrutura de cada repositório; detectar e resolver conflitos de semântica; e realizar a integração dos dados a partir de consultas a cada repositório de interesse. Através da X-ARC o usuário pode adiar a fase de detecção e resolução de conflitos de semântica realizando consultas diretas a um ou mais repositórios utilizando para isso termos de domínios associados a repositórios de dados.

A partir de uma consulta enviada para a X-ARC e através dos metadados mantidos pelo Gerente de Metadados o processador de consultas da arquitetura é capaz de decompor a consulta em várias subconsultas que são enviadas para os repositórios apropriados. O serviço de processamento de consultas da arquitetura

controla a execução de cada uma das subconsultas enviadas e aguarda o retorno do serviço de Acesso aos Dados para cada um dos resultados para depois agrupá-los em um resultado único e retorná-lo para a aplicação (cliente) que disparou a consulta.

Esta é uma tarefa do mediador que utiliza os metadados mantidos pela arquitetura para resolver como, quando e o que executar da consulta. O mediador recebe a consulta que engloba repositórios espaciais e não-espaciais e aplica um conjunto de regras para identificar que parte da consulta deve ser atendida por um ou mais repositórios, desmembra a consulta em um ou mais subconsultas e envia as subconsultas para o mediador ou tradutor que deverá tratar a consulta. O algoritmo em pseudo-código para desmembrar uma consulta em subconsultas e determinar se a consulta pode ser atendida pela arquitetura é apresentado na Figura 4.17.

```
Início-Algoritmo
Se houver caractere "<" após cláusula From então;
    Captura Termo de Domínio indicado na cláusula From;
    Se Termo de Domínio existe no Gerente de Metadados então;
        // consulta pode ser atendida e desmembrada
        Solicita ao Gerente de Metadados lista de URLs equivalentes ao Termo de Domínio;
        Para cada URL
            reescreve a consulta substituindo o Termo de Domínio na cláusula From pela URL;
            Executa consulta reescrita;
    Senão
        //consulta não pode ser atendida
        Retorna Erro;
Senão
    Captura URL indicada na cláusula From;
    Se URL existe no Gerente de Metadados então;
        // consulta pode ser atendida
        Executa a consulta;
    Senão
        //consulta não pode ser atendida
        Retorna Erro;
Fim-Algoritmo
```

Figura 4.17 – Algoritmo de Re-Escrita de Consulta

A arquitetura é capaz de unificar os resultados provenientes de repositórios contendo estruturas de dados e esquemas variados graças à utilização da XML como linguagem padrão de intercâmbio de dados. A Figura 4.18 apresenta os passos envolvidos no processamento de consulta da arquitetura.

As consultas são traduzidas para as linguagens de consulta dos repositórios locais e distribuídas aos mesmos para: *i)* acessar as múltiplas formas dos dados armazenados sob diversos ambientes; *ii)* recuperar os dados desejados; *iii)* integrar o resultado da consulta; e *iv)* apresentar os resultados, suportando visualização gráfica destes na forma de mapa e/ou tabela.

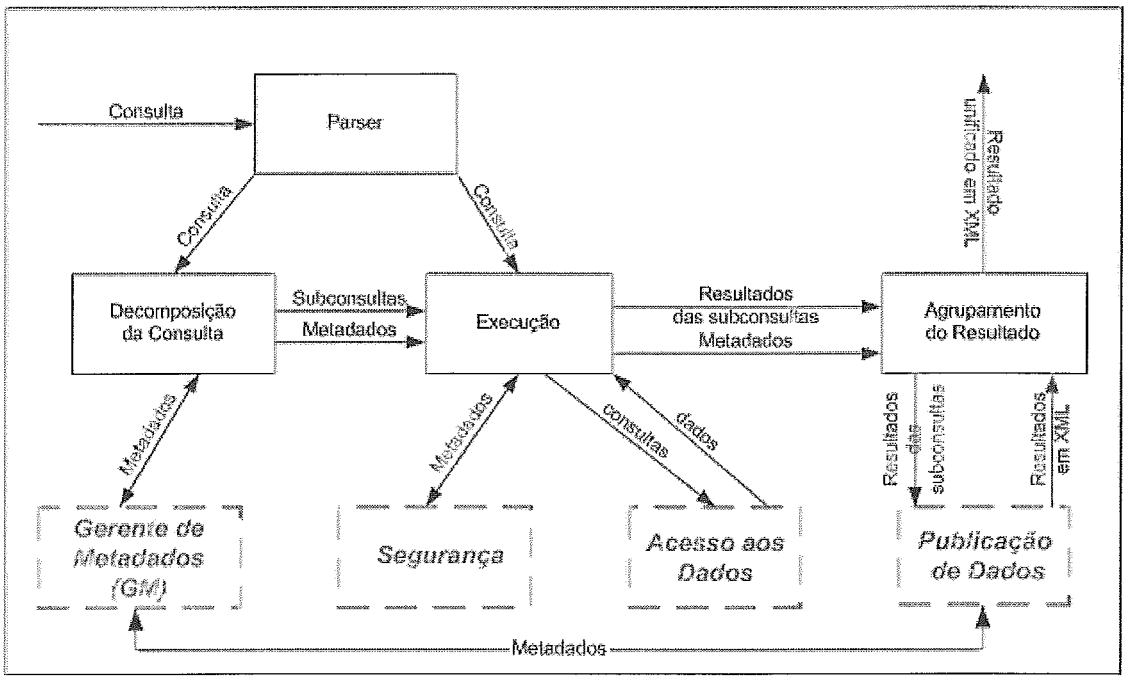


Figura 4.18 – Passos do Processamento de Consulta

No Passo 1 a consulta é verificada pelo serviço de Processamento de Consultas para determinar se trata-se de uma consulta direcionada a um repositório específico ou se a mesma deve ser divida em subconsultas e estas enviadas aos repositórios adequados. Caso deva ser subdividida o Passo 2 a seguir é realizado, caso contrário o Passo 3 de execução da consulta deve ser realizado.

No Passo 2 a consulta inicial é re-escrita em tantas subconsultas, quantas forem necessárias, para serem enviadas ao passo seguinte que irá enviar as subconsultas para os repositórios específicos. Neste passo o GM é consultado para se obter as fontes de dados que possuem a mesma classificação de termo de domínio utilizado na consulta em questão. A utilização de termos de domínio na X-ARC apoia a localização de possíveis fontes de dados uma vez que permite que repositórios inicialmente desconexos sejam relacionados a partir de termos de domínios atribuídos às suas fontes.

O Passo 3 é responsável pela execução das subconsultas. Este passo é executado pelo serviço de Acesso aos Dados da arquitetura. Cada subconsulta é enviada para o tradutor do repositório que terá como tarefa mapear a consulta expressa no modelo de dados comum da arquitetura para o modelo de dados da fonte de dados. Após a execução das subconsultas o resultado de cada uma delas deve ser unido para representar um único resultado retornado pela arquitetura. Este é o papel desempenhado pelo Passo 4 do processamento de consultas.

No Passo 4 todas os resultados das subconsultas são unidos em um único resultado para ser retornado para a aplicação que enviou a consulta inicial. Embora a maioria das fontes apresente estruturas diferentes, os dados são unificados pela X-ARC. O recurso utilizado para alcançar esta unificação de resultados é o uso da XML como padrão de intercâmbio de dados da arquitetura, onde a partir dos resultados retornados por cada subconsulta um documento XML "bem formado" é construído pelo Serviço de Publicação de Dados anexando cada resultado parcial das subconsultas no documento XML formando uma coleção de dados única. Maiores detalhes sobre a geração do documento XML e o Serviço de Publicação de Dados serão apresentados na próxima seção.

4.10 Publicação de Dados

A publicação de dados da arquitetura desempenha um papel importante na unificação dos dados disponibilizados pelos participantes. Uma vez que a arquitetura provê acesso a dados espaciais e não-espaciais de uma forma integrada, a disponibilização destes dados para os usuários da arquitetura torna-se de fundamental importância.

Tendo em vista que a integração de integração de dados convencionais e não-convencionais (SOUZA, 1986, STRAUCH, 1998) é um processo complexo e que muitas vezes necessita de intervenção humana, a nossa proposta ao invés de procurar resolver todos os conflitos que podem ocorrer durante o processo de integração de dados espaciais e não-espaciais, tais como: diferenças de nomes (tanto sinônimos como homônimos), diferenças de formato, diferenças estruturais, conflitos de dados e conflitos de contexto espacial, utiliza a linguagem XML para disponibilizar estes dados de forma unificada, deixando a critério do cliente que solicitou o acesso aos dados a decisão de que fontes de dados utilizar e de que forma combiná-las para utilizá-las.

Assim sendo, a arquitetura provê uma forma padrão de acesso aos dados, tanto espaciais como não-espaciais, de forma que suas heterogeneidades (formato, representação, estrutura, etc.) não impeçam sua localização e utilização pelos usuários.

A arquitetura X-ARC emprega a linguagem XML para apoiar a publicação de dados dos diversos repositórios de dados envolvidos e este é um dos fatores que diferenciam a arquitetura das outras soluções. Porém, não é o uso da linguagem XML que a diferencia, mas como é utilizada na arquitetura. Pois ao invés de mapear todos

os repositórios envolvidos para uma representação em XML e, consequentemente, perder o poder de processamento de consulta específico de cada um dos repositórios, a X-ARC mantém os dados onde estes se encontram e apenas os disponibilizam em formato XML, facilitando o processo de publicação dos repositórios envolvidos.

A X-ARC possui o Serviço de Publicação de Dados que é responsável pela publicação dos dados disponibilizados pela arquitetura, o qual através da estrutura do resultado proveniente de cada acesso a um repositório, constrói um documento XML, acrescentando os metadados associados ao(s) repositório(s) e ao(s) dado(s) consultado(s). O documento XML é criado de duas formas, de acordo com a natureza do dado acessado. Para dados descritivos, o conteúdo (tuplas) retornado pelo serviço de Acesso aos Dados é inserido no documento XML "bem formado" criado. No caso dos dados espaciais, estes serão publicados como uma imagem. Assim sendo, ao invés de inserir as tuplas dos dados espaciais no documento, uma referência à imagem que os representa é incluída no documento XML.

Este serviço é responsável por transformar os dados para uma representação que combina/integra dados recuperados de repositórios heterogêneos. Ele realiza um pós-processamento que reformata os dados recuperados a serem apresentados ao usuário.

Uma vez que um documento XML é criado para cada consulta enviada para a arquitetura, não se pode falar que o serviço de publicação utilize um DTD único para publicar os dados. Entretanto, pode-se considerar que existe um DTD inicial que é expandido à medida que os resultados das subconsultas são incorporados em um único documento.

A Figura 4.19 apresenta o padrão de DTD proposto para a X-ARC na construção dos documentos XML que encapsulam os dados publicados pela arquitetura.

```
// resultado é composto por zero ou mais sites
<!ELEMENT RESULTADO (SITE)*>
// um site é composto por zero ou mais linhas
<!ELEMENT SITE (LINHA)*>
// uma linha é composto por zero ou mais colunas
<!ELEMENT LINHA (COLUNA)*>
// uma coluna é um conjunto de caracteres
<!ELEMENT COLUNA (#PCDATA)>
<!ATTLIST SITE
        url CDATA #REQUIRED
>
```

Figura 4.19 – Padrão de DTD para Publicação dos Dados na X-ARC

Um exemplo de como o Serviço de Publicação de Dados unifica os dados publicados pela arquitetura é apresentado a seguir. As Figuras 4.20 e 4.21 apresentam os resultados intermediários de uma consulta enviada à arquitetura. Na Figura 4.20 é possível observar a substituição do elemento "Coluna" do DTD da Figura 4.19 pelos elementos "Data" e "Velocidade" que foram incluídos pelo Serviço de Publicação de Dados durante a geração do documento. Na Figura 4.21 observa-se o elemento "Figura" em substituição ao elemento "Coluna" do DTD, também gerado pelo serviço durante a geração do documento. O resultado da Figura 4.20 é um dado não-espacial, enquanto que o da Figura 4.21 é um dado espacial.

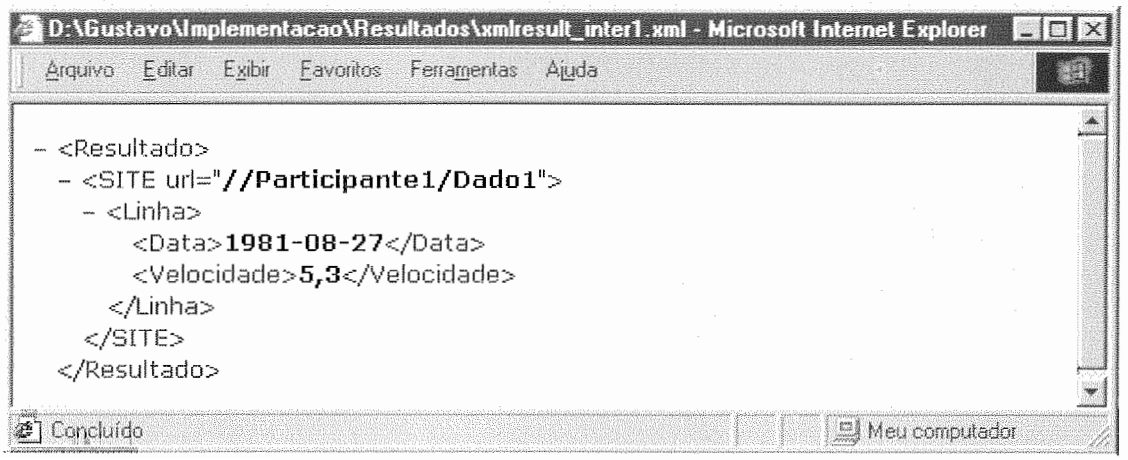


Figura 4.20 – Resultado Intermediário 1

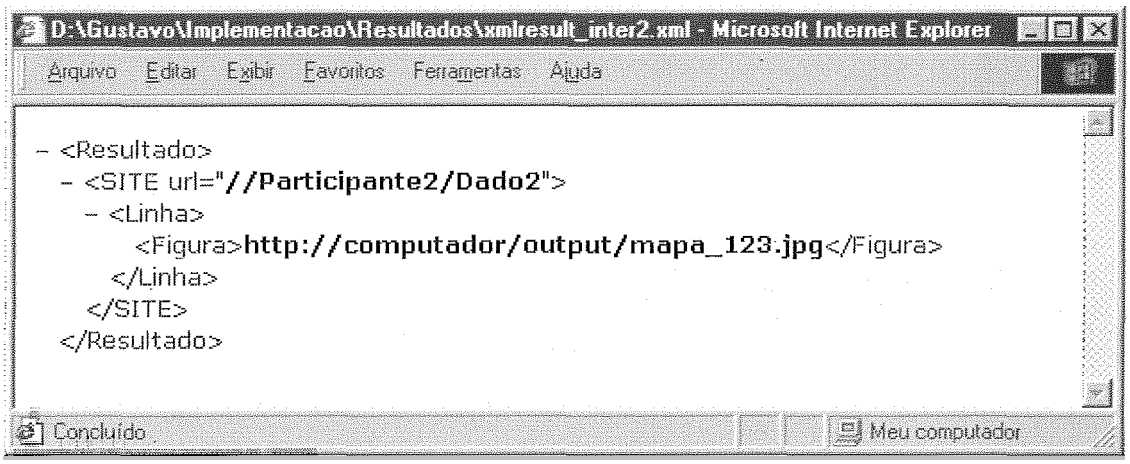


Figura 4.21 – Resultado Intermediário 2

A partir dos resultados intermediários, o Serviço de Publicação de Dados unifica os resultados em um único (Figura 4.22), identificando os repositórios de dados, mantendo os dados retornados pela arquitetura e preservando a estrutura dos dados de cada resultado intermediário.

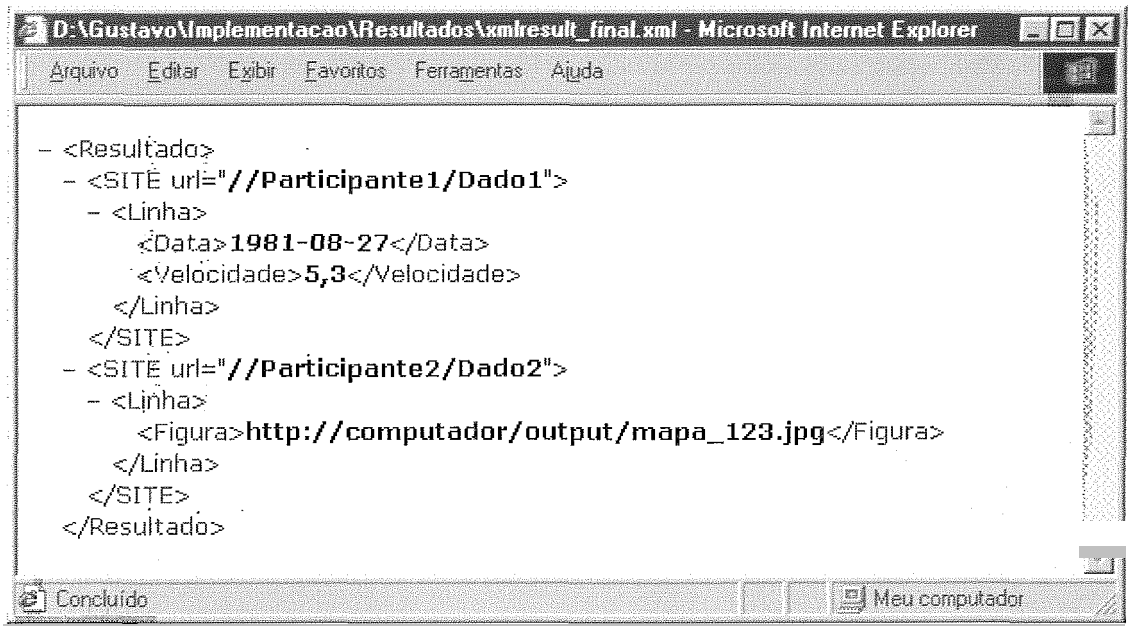


Figura 4.22 – Resultado Final gerado pelo Serviço de Publicação de Dados

Uma vantagem do uso da XML na exibição dos resultados da consulta é a facilidade de apresentação do resultado ao usuário, permitindo que interprete o resultado de acordo com a semântica da informação contida no documento.

5 ESTUDO DE CASO DA ARQUITETURA

Este capítulo tem por objetivo apresentar o estudo de caso no qual um protótipo da arquitetura X-ARC foi implementado para validar as funcionalidades da mesma. Este estudo foi realizado no âmbito do ambiente AGROMET que fornece um ambiente de trabalho cooperativo para usuários da área de Agrometeorologia durante o processo de tomada de decisão.

O estudo de caso é realizado a partir das necessidades surgidas durante a execução do projeto de preservação da Mata Atlântica, desenvolvido pela EMBRAPA Solos, que durante sua execução necessitou localizar, disponibilizar e acessar dados de Agrometeorologia heterogêneos e distribuídos.

As próximas seções enfocarão uma descrição do ambiente AGROMET, o detalhamento do estudo de caso, a caracterização do dados envolvidos e o protótipo desenvolvido. Durante as seções será exposto o problema dos usuários no acesso aos dados e a necessidade por uma solução que seja capaz de integrá-los dado sua natureza heterogênea; e como a X-ARC pode contribuir para o desenvolvimento das atividades dos usuários de dados ambientais.

5.1 AGROMET

O Projeto AGROMET (SOUZA, STRAUCH, PINTO et al., 2002) oferece um ambiente de trabalho cooperativo, comum, flexível e de uso intuitivo, no qual os membros de um grupo podem estar geograficamente distribuídos em ambientes heterogêneos, interagindo na tomada de decisões. A possibilidade de aplicação de tecnologias de *groupware* para dar suporte às decisões de cunho ambiental com a

participação colaborativa de diversos grupos de interesses distintos constitui a base do projeto.

Neste contexto, o projeto procura integrar atividades de pesquisadores, cientistas sócio-econômicos, representantes das comunidades locais, políticos, produtores, e diversos outros participantes que formam equipes multidisciplinares e heterogêneas envolvidas no processo de tomada de decisão sobre diversos assuntos tais como preservação ambiental, planejamento de crescimento urbano, etc.

O Projeto AGROMET é composto por uma arquitetura em camadas, onde as camadas superiores fornecem ferramentas de apoio à colaboração, cooperação, gerenciamento de *workflow* científico, gestão de documentos e conteúdo; e a camada inferior fornece acesso a dados que são utilizados pelos usuários envolvidos na tomada de decisão.

A arquitetura X-ARC compõe a camada de acesso aos dados da arquitetura AGROMET sendo responsável pela integração e compartilhamento dos repositórios heterogêneos utilizando para isso mediação. Desta forma, a X-ARC provê serviços de acesso a dados para as camadas superiores da arquitetura do projeto AGROMET.

5.2 Detalhamento do Estudo de Caso

Como foi citado no capítulo 1, os usuários que participam dos processos de tomada de decisão na área ambiental são multidisciplinares, ou seja, provêem de várias áreas de conhecimento tais como: meteorologia, engenharia, sociologia, pedologia, cartografia, etc. Além disso, o processo de análise no qual eles estão envolvidos utiliza um conjunto variado de dados, porém o custo de obtenção destes dados é muito alto, resultando em uma troca de dados informal entre as instituições provedoras de dados e os usuários/instituições consumidoras de dados.

O problema reside justamente na ausência de padrões que determinam a estrutura, formato, armazenamento, etc. destes dados, implicando em uma babel de formatos, estruturas e formas de armazenamento proprietários determinados pela indústria e por cada organização que coleta, processa e disponibiliza os dados.

Estes dados encontram-se espalhados, fisicamente, em diversos locais, armazenados em diversos formatos e por sistemas diferentes, porém os pesquisadores necessitam acessá-los.

Além disso, em alguns casos não há o prévio conhecimento de que os dados encontram-se disponíveis para o trabalho, portanto é necessário que seja fornecido uma ferramenta que possibilite a busca por dados de acordo com determinadas características e que a estrutura destes dados seja informada aos usuários, permitindo que decidam sua utilização ou não. Tal ferramenta deve ser capaz de localizar repositórios de dados relevantes para o desenvolvimento das atividades e disponibilizar um acesso integrado aos mesmos. Desta forma, independente do formato ou localização dos dados, eles serão manipulados sempre da mesma forma.

A arquitetura X-ARC foi concebida para fornecer uma visão unificada dos dados utilizados pelos usuários da área ambiental, mesmo que estes dados apresentem

as heterogeneidades citadas anteriormente e um protótipo desta arquitetura foi construído a fim de validá-la.

5.3 Caracterização dos Dados

No estudo de caso em questão, a X-ARC é utilizada para prover acesso a:

- dados produzidos por uma estação meteorológica, em um formato proprietário determinado pela empresa fabricante, no caso em questão trata-se de um arquivo texto delimitado;
- uma base de dados contendo séries históricas de temperatura, pressão e precipitação, mantida por uma grupo de pesquisadores de uma organização;
- um conjunto de planilhas eletrônicas que armazenam informações sobre vento, tais como direção, velocidade, etc.;
- uma base de mapas da região do projeto, gerada através de dados coletados pelos diversos pesquisadores envolvidos no projeto e armazenada em um Sistema de Informação Geográfica.

A seguir será fornecida uma breve descrição a respeito das características específicas dos repositórios de dados que participam do estudo de caso.

5.3.1 Não-Espaciais

Os dados não-espaciais envolvidos no estudo de caso em questão são dados climáticos provenientes de instituições públicas ou privadas especializadas na coleta e processamento destes dados. A própria instituição, EMBRAPA Solos, gera e processa parte destes dados.

Uma característica importante que vale a pena ressaltar sobre estes dados é a heterogeneidade do formato em que se apresentam. Entre os diversos formatos de armazenamento existentes, podemos ressaltar: o formato texto tabular e o formato de planilha eletrônica, por exemplo Excel.

Outro detalhe importante com relação a estes dados é a variedade do conjunto de informações que compõe estas fontes de dados, ou seja, cada instituição fornece um conjunto variado de informações. Além disso, não existe um padrão que determine a posição em que estes conjuntos devem ser enviados, o que resulta em alterações na ordem dos conjuntos para uma mesma instituição qui sá entre instituições.

Completando as questões relativas à heterogeneidade dos dados, pode se acrescentar a granularidade da informação, visto que os dados podem ser:

- de observações ao longo do dia (manuais ou automáticas);
- diários;
- mensais; ou
- anuais.

Cabe ressaltar a ausência de informações completas, o que resulta na utilização de médias para o preenchimento das lacunas e na necessidade da adoção de um grau de confiabilidade e qualidade do dado.

A seguir, é apresentado nas Figuras 5.1, 5.2 e 5.3, uma amostra dos dados envolvidos no estudo de caso. Nas Figuras 5.1 e 5.2 é possível observar a estrutura dos dados envolvidos, enquanto que na Figura 5.3 tem-se uma visualização dos dados espaciais.

Estacao1.txt - Notepad

File Edit Format Help

LATITUDE	LONGITUDE	DATA	HORA	TEMPERATURA	PRECIPITAÇÃO
-0,80	-62,90	27-08-2001	17	38	35,1
-0,80	-62,90	27-08-2001	18	34,1	6,7
-0,80	-62,90	27-08-2001	19	36,9	6,8
-0,80	-62,90	27-08-2001	20	35,3	7,2
-0,80	-62,90	27-08-2001	21	33	0
-0,80	-62,90	27-08-2001	22	33,1	0
-0,80	-62,90	27-08-2001	23	31,4	0

Figura 5.1 – Exemplo dos dados gerados pela estação

Microsoft Access - [SERIE HISTORICA : Table]

File Edit View Insert Format Records Tools Window Help

ANO	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1961	305,2	164,9	473,3	458,7	400	288,3	410,7	222	249,5	365,1	322,9	363,6
1962	391,8	289,2	253	328,2	527,2	434,6	245,6	157,2	188,9	271,7	171	143,6
1963	238	155,7	386,3	429,3	400,3	220,3	287,3	268,3	225	153,1	239	280,9
1964	148,8	324,5	332,6	311,2	341,2	371	369,1	238,9	374,6	321	267,5	327,5
1965	236,7	260,7	406,6	264,7	340,4	355,9	235,9	189,7	329,2	228,9	172	276,4
1966	310,6	249,6	281,6	266,9	349	403,9	379,3	180,9	234,3	144,5	102,5	128,8
1967	153,2	349,2	277,2	255,7	504,7	274,5	414,9	243,3	209,3	320,2	254,6	103
1968	372	229,3	385,7	553,6	261,3	460	344,5	277,5	314,6	105	269,7	298,5
1969											86,6	224,3

Record: 1 of 58

Datasheet View CAPS NUM

Figura 5.2 – Exemplo dos dados da série histórica

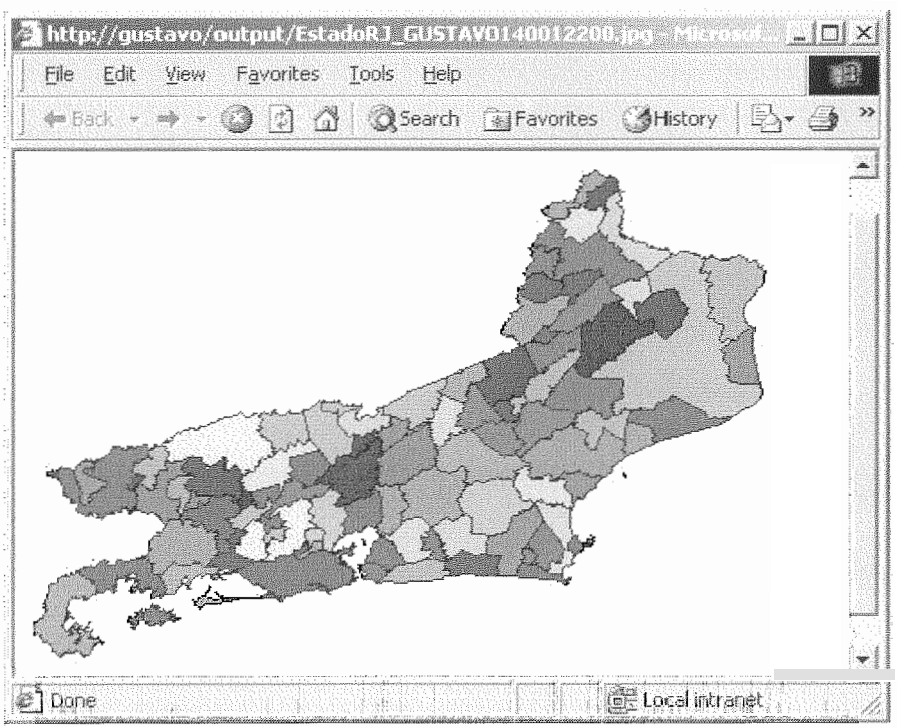


Figura 5.3 – Exemplo dos dados espaciais

5.3.2 Espaciais

Os dados espaciais que constituem o estudo de caso, assim como os dados não-espaciais, são dados obtidos de outras empresas ou gerados pela EMBRAPA Solos a partir de levantamentos realizados em campo.

Estes dados encontram-se em formato DXF, ArcInfo Coverage, TIFF, Shape File, etc. Todos os dados que compõem uma ou mais camadas de informação relativas ao projeto de decisão são tratados antes que sejam inseridos na base cartográfica. O tratamento destes dados é feito utilizando o sistema ArcInfo. Após o tratamento estes dados são registrados no ArcIMS para consulta pelos usuários envolvidos no projeto. Embora os dados estejam disponíveis, não existe um catálogo geral sobre estes dados.

Alguns exemplos de dados espaciais são: altimetria, hidrografia, pedologia, cobertura vegetal, etc. Cada um destes constituem uma ou mais camadas de informação que são utilizadas pelos tomadores de decisão e portanto precisam estar disponíveis durante o processo de decisão. Todavia, pode ocorrer a existência de dois ou mais repositórios que possuem dados de altimetria, hidrografia, etc. disponíveis para serem utilizados. A localização destas alternativas é importante, pois fornece informações que auxiliam a decisão dos usuários, seja apoiando, reforçando ou complementando suas observações.

5.4 Implementação do Protótipo X-ARC

O protótipo da arquitetura X-ARC foi desenvolvido com o objetivo de avaliar a viabilidade de aplicação da proposta em uma situação real em que os usuários necessitam da interoperabilidade dos dados.

O protótipo foi concebido para executar uma instância X-ARC que pode ser configurada com o perfil de **Gerente** ou **Participante** (tarefa realizada pelo Usuário Administrador) e ao mesmo tempo ser a aplicação utilizada pelos usuários da arquitetura que disponibilizam e procuram dados (Usuário Produtor e Usuário Consumidor) através do uso dos serviços da arquitetura como o Assistente de Publicação de Dados e o Serviço de Processamento de Consultas.

O protótipo da X-ARC foi implementado segundo o paradigma de orientação a objetos utilizando como linguagem de implementação **JAVA** e como ambiente de desenvolvimento o **JBuilder 6**, da Borland. A linguagem JAVA foi escolhida pois trata-se de uma linguagem de implementação que não necessita e não se restringe a nenhuma plataforma específica de *hardware* ou *software* para sua execução. Desta forma, o protótipo inicial da arquitetura pode ser executado em qualquer recurso computacional, desde que exista uma máquina virtual java (JVM) disponível.

O banco de dados utilizado para armazenar os metadados armazenados e mantidos pela arquitetura foi o banco de dados relacional **ACCESS** da Microsoft. O acesso ao banco da arquitetura é realizado através de uma ponte **JDBC-ODBC**. O banco ACCESS foi escolhido a fim de evitar gastos desnecessários de tempo com a configuração de bancos relacionais mais robustos como o ORACLE e o SQLServer.

O projeto do protótipo foi elaborado baseado em técnicas de padrões de projeto (Design Patterns) (GAMA, HELM, JOHNSON et al., 2000) visando implementar o protótipo de forma a permitir: a reutilização de suas classes Java pelos outros

membros dos projetos SPeCS, CoAGRI e AGROMET que desenvolvem protótipos complementares à X-ARC; e a flexibilidade na implementação do protótipo reduzindo o impacto das alterações na implementação futura da arquitetura. As principais funções do protótipo encontram-se relatadas nos diagramas de Casos de Uso presentes no Apêndice A.

As principais classes Java utilizadas para a implementação do protótipo encontram-se representadas pelo diagrama de classes no Apêndice B, enquanto que uma listagem do código do Serviço de Processamento de Consulta é apresentado no Apêndice D.

Durante a análise do estudo de caso a ser apresentado na dissertação foi identificado a necessidade da extensão dos tradutores do Le Select para dados armazenados em formato Excel, pois não se dispunha de um tradutor específico para o formato. Para desenvolver o tradutor específico para o Excel foi utilizada uma biblioteca *OpenSource* implementada em Java denominada **"ExcelRead"** juntamente com a extensão das classes básicas do pacote **WrapperInterface** do Le Select que são responsáveis pela criação de tradutores para os repositórios. O diagrama contendo as classes implementadas para o novo tradutor encontram-se no Apêndice C. O novo tradutor implementado foi acrescentado ao servidor Le Select, durante sua inicialização, como um arquivo JAR, que foi gerado com todas as classes estendidas do Le Select juntamente com a API da biblioteca ExcelRead.

A comunicação entre as instâncias distribuídas da arquitetura (gerentes e participantes) ocorre utilizando a solução *Remote Method Invocation* (RMI) da linguagem Java. Enquanto que a comunicação entre a arquitetura e os sistemas intermediários Le Select e ArcIMS é realizada através dos tradutores X-SELECT e X-MAP, respectivamente.

A comunicação entre o X-SELECT e o servidor Le Select é realizada através de uma conexão estabelecida com o servidor Le Select através de uma API (arquivo JAR), fornecida juntamente com a instalação do Le Select, a qual fornece um *driver* de conexão JDBC e as classes para a manipulação dos resultados e envio de consultas ao servidor.

A comunicação entre o X-MAP e o servidor ArcIMS é realizada através de uma conexão, por ***sockets***, estabelecida com o servidor de páginas html (servidor Web) que recebe o pedido em ArcXML. Em seguida, o servidor html repassa o pedido para o servidor ArcIMS que processa o pedido e retorna o resultado em ArcXML para o servidor html repassá-lo para o tradutor X-MAP. A conexão é estabelecida através de uma API (arquivo JAR), que fornece a classe para o envio de pedidos e o recebimento dos resultados. As classes dos tradutores X-SELECT e X-MAP e seus relacionamentos encontram-se no Apêndice B.

5.5 Protótipo X-ARC

Partindo do pressuposto de que os repositórios de dados envolvidos constituem-se de: o arquivo gerado pela estação agrometeorológica, a base de dados de série histórica, as planilhas Excel e os mapas armazenados no SIG, propõe-se a seguinte configuração da arquitetura para o estudo de caso: uma instância com perfil de Gerente e duas instâncias com perfis de Participante, sendo que uma estará disponibilizando dados não-espaciais e a outra estará disponibilizando dados espaciais (mapas).

Para o acesso aos dados armazenados em planilha Excel, o LeSelect não dispunha de um tradutor padrão, portanto foi necessário desenvolver um tradutor específico para planilhas Excel e acoplá-lo ao Le Select. Esta capacidade de adaptação justifica uma das razões que nos levaram a utilização do Le Select na arquitetura. Para os outros repositórios de dados utilizamos tradutores já existentes, assim o próximo passo para disponibilizar os repositórios é registrar cada um deles.

A disponibilização dos dados envolvidos em cada participante é realizado utilizando-se o Assistente de Publicação de Dados (Figura 5.5). O assistente acompanha o usuário no processo de disponibilização dos dados, capturando os metadados, gerando um tradutor e associando, com o auxílio do usuário, um termo de domínio para cada um dos dados.

Inicialmente o Usuário Publicador, que está disponibilizando o dado, informa o tipo de dado a ser disponibilizado (espacial ou não-espacial). A Figura 5.4 ilustra o registro de um dado não-espacial, onde o usuário atribui um nome para o dado sendo disponibilizado, informa como este dado é armazenado, fornece o banco de dados que o armazena (caso seja armazenado em um banco de dados) ou o formato do arquivo

(caso seja armazenado em arquivo) e atribui um termo de domínio que melhor classifica o dado.

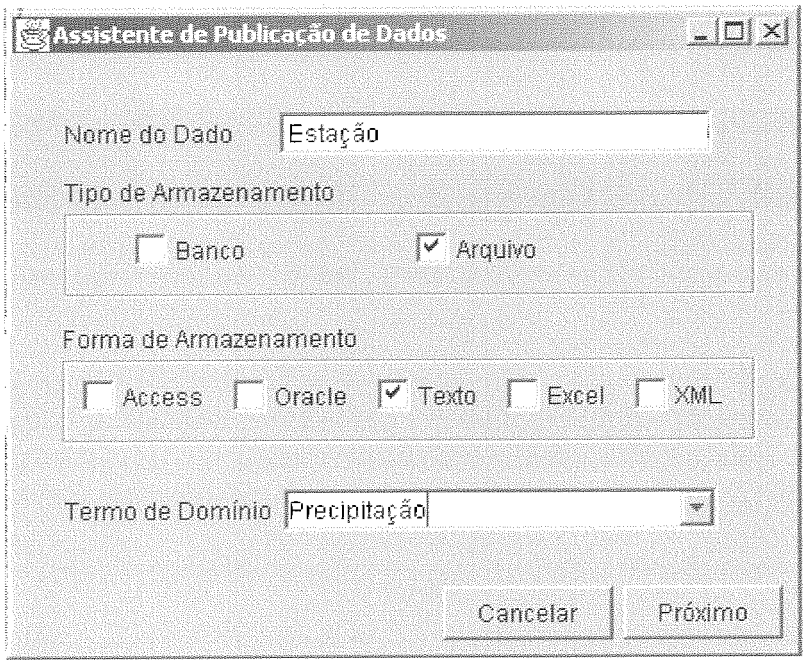


Figura 5.4 – Assistente de Publicação de Dados

O próximo passo no assistente é informar a localização do dado sendo disponibilizado, fornecendo dados como IP da máquina que armazena o dado, localização do dado na máquina, nome de usuário e senha para conectar (caso seja necessário), etc. Em seguida, o usuário deve indicar o nível de acesso que os outros usuários da arquitetura terão ao acessar o dado e os metadados que descrevem o dado sendo publicado. O nível de acesso é atribuído por usuário, mas ainda é mantido um nível de acesso padrão a ser utilizado pela arquitetura caso o usuário que acessa o dado não esteja cadastrado na arquitetura.

O último passo na publicação dos dados realizado pelo usuário é fornecer a estrutura do dado que está sendo disponibilizado, como nome do atributo, tipo, tamanho, etc.

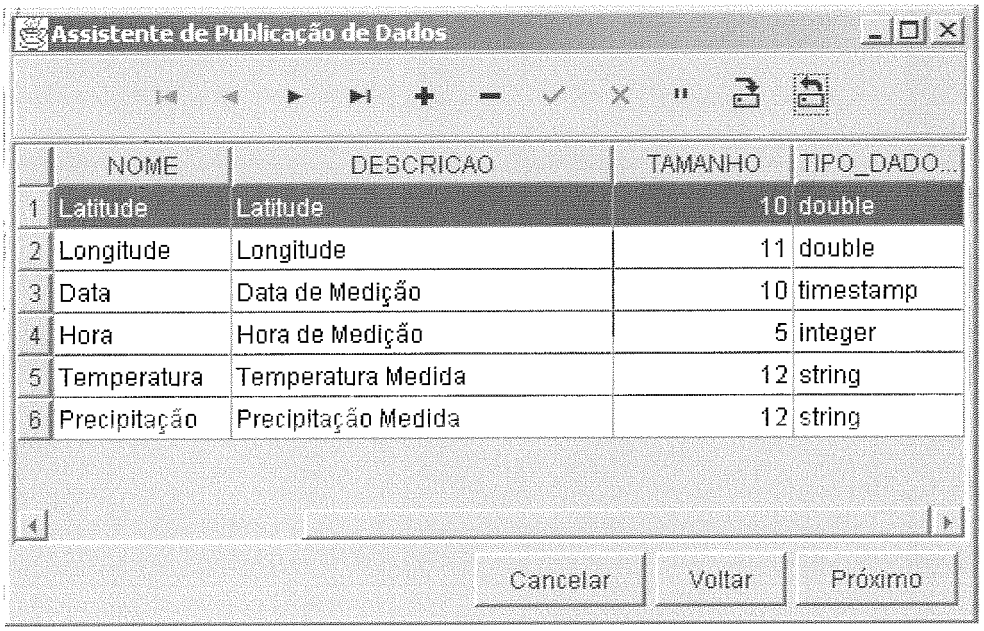


Figura 5.5 – Assistente de Publicação de Dados (estrutura do dado)

Para cada dado não-espacial publicado pelo assistente, obtém-se ao final de cada registro um arquivo com a definição do tradutor para o dado. Por exemplo, a Figura 5.6 mostra o conteúdo do arquivo com a definição do tradutor para os dados gerados pela Estação Meteorológica que são armazenados em um arquivo texto delimitado.

```
<Wrapper WrapperClass="LeSelect.Wrappers.Text.TableWrapperFactory">
        <Parameters>
        <Table name="Estacao" file="E:/Dados/Caso/Estacao1.txt" >
                <Column name="Latitude" type="double" size="10"/>
                <Column name="Longitude" type="double" size="11"/>
                <Column name="Data" type="timestamp" size="10"/>
                <Column name="Hora" type="integer" size="5"/>
                <Column name="Temperatura" type="string" size="12"/>
                <Column name="Precipitação" type="string" size="12"/>
        </Table>
        </Parameters>
</Wrapper>
```

Figura 5.6 – Arquivo de Definição do Tradutor para Estação

É importante lembrar que durante o registro dos repositórios, a estrutura dos dados e o termo de domínio informados pelo usuário são coletados pelo Gerente de Metadados e, enviados ao gerente para compor os metadados globais da arquitetura.

Uma vez que os dados encontram-se disponibilizados pela X-ARC, a localização e o acesso a dados torna-se fácil e uniforme, independente do formato e local em que se encontram. Para tal, um usuário que necessita de dados relacionados ao termo "precipitação" pode utilizar a X-ARC para localizar e recuperar os dados. A Figura 5.7 exibe a interface do protótipo que o usuário utiliza para localizar os dados que necessita. Nesta interface o usuário escolhe o termo de domínio de seu interesse e em seguida a consulta é enviada para a X-ARC. Para o exemplo da figura, a seguinte consulta é enviada para a arquitetura: *"SELECT * FROM <PRECIPITAÇÃO>"*.

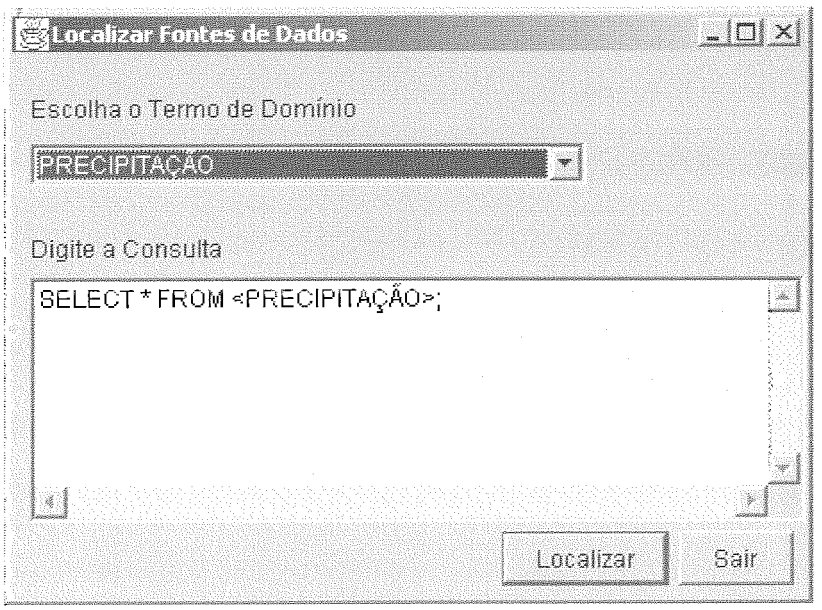


Figura 5.7 – Consulta a dados por termo de domínio "Precipitação"

A consulta é recebida pelo Processador de Consulta (PC) do Gerente o qual verifica se a consulta direciona-se a um repositório de dados específico ou não. Neste caso, a consulta utiliza um termo de domínio para englobar o maior número de

repositórios possíveis. Assim, o PC re-escreve a consulta baseado nos metadados globais do gerente, então reenvia para cada participante, que classifica seu dado segundo o termo escolhido, as consultas re-escritas (sub-consultas) para os repositórios existentes. O resultado da execução de cada sub-consulta é retornado para o gerente que une os resultados em um único e o publica através do serviço de Publicação de Dados.

O resultado unificado da consulta é apresentado na Figura 5.8. Conforme pode ser observado, os dados da estação meteorológica e da base de dados em Access são publicados e disponibilizados virtualmente para os usuários da Embrapa Solos de forma uniforme.

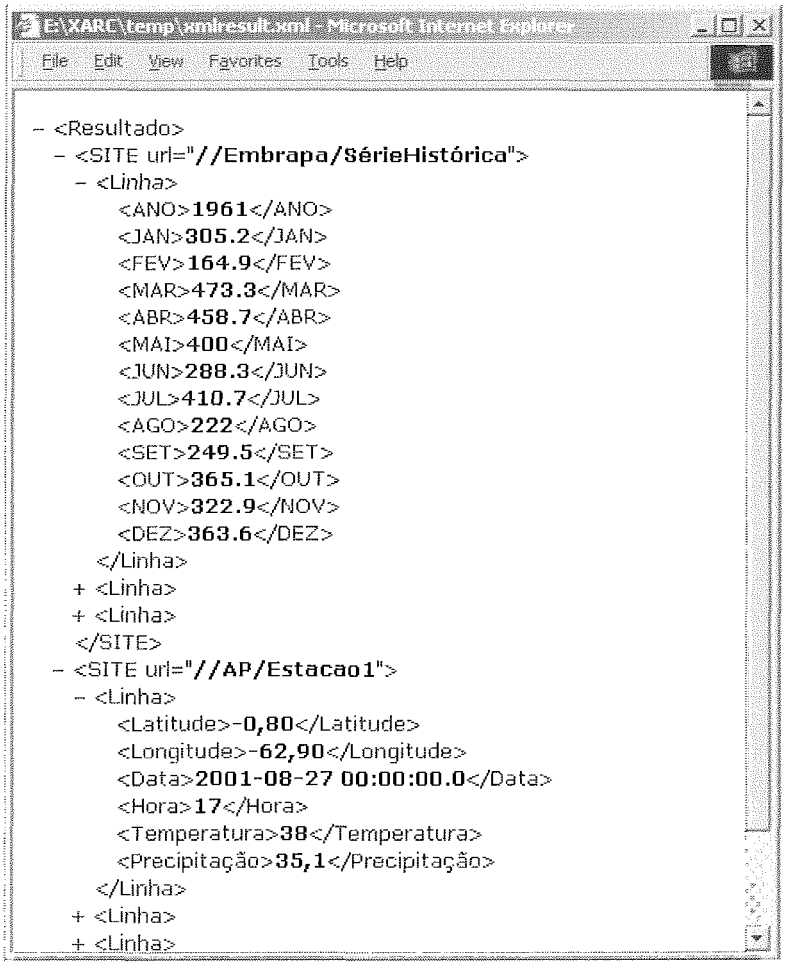


Figura 5.8 – Resultado unificado (XML) da consulta a dados de Precipitação

6 CONCLUSÃO

Uma vez que o compartilhamento de dados e serviços tem se tornado cada vez mais determinante para o desempenho das tarefas das organizações e que a cada momento novos repositórios de dados são criados e disponibilizados a partir da Internet, torna-se imprescindível disponibilizar soluções que permitam o gerenciamento destes dados e serviços com relação à localização, acesso e permissão de uso dos mesmos pelos usuários.

Baseado no conceito introduzido por TOMASIC & SIMON (1997) que atribui papéis ora de consumidores ora de produtores de dados aos usuários e no crescimento incessante da Internet, torna-se evidente a necessidade pela interoperabilidade de dados.

A partir deste cenário, várias propostas surgiram para solucionar os problemas de integração existentes atualmente, sendo uma delas a mediação de dados, segundo a qual a arquitetura X-ARC – e**X**tensible **Arc**hitecture, apresentada neste trabalho, está baseada.

Esta dissertação está inserida no contexto de três projetos de pesquisa do Programa de Engenharia de Sistemas e Computação da COPPE/UFRJ, a saber: Ambiente de Apoio à Decisão Cooperativa para Agricultura de Precisão (**CoAgri**), Sistema de Suporte à Decisão Espacial Colaborativo (**SPeCS**) e Agrometeorológico (**AGROMET**). Nesses projetos, a X-ARC compõe a camada de integração de dados e tem como função prover serviços de acesso a dados para as camadas superiores das arquiteturas dos projetos.

A arquitetura X-ARC tem como objetivo disponibilizar o acesso a dados espaciais e não-espaciais heterogêneos e distribuídos de forma que a localização e a heterogeneidade dos repositórios seja transparente para o cliente (usuário/aplicação).

A linguagem XML é utilizada como um padrão de intercâmbio de dados, uma vez que sua estrutura é flexível e capaz de se adaptar aos diferentes formatos dos dados. Assim, os dados dos diversos repositórios envolvidos são acessados uniformemente pelos usuários, reduzindo a heterogeneidade de formato dos dados.

A X-ARC encapsula dois sistemas para a execução de seus serviços e apoio à publicação dos dados: o Le Select (LESELECT, 2002), *midleware* do INRIA que manipula dados estruturados e semi-estruturados; e o ArcIMS (ARCIMS, 2001) que manipula dados georreferenciados.

Termos de domínio são utilizados para auxiliar na associação dos dados heterogêneos aos seus domínios de aplicação específicos e um conjunto de metadados é mantido pela arquitetura para apoiar seu gerenciamento e o processamento das consultas.

6.1 Análise das Contribuições

A X-ARC é uma arquitetura flexível e escalável, pois novos repositórios de dados podem ser acrescentados ou retirados da arquitetura, sem que haja esforço por parte dos usuários e sem aumento na complexidade de sua gerência.

Para a área ambiental, domínio de aplicação da arquitetura, a utilização de um esquema global para o mediador foi substituído pelo uso de termos de domínio que classificam os conjuntos de dados e permitem uma simplificação do processo de mediação, visto que a tecnologia de mediação aplicada na arquitetura foi estendida para permitir a publicação dos dados através de coleções de dados, conforme explicado na seção 4.3.

A arquitetura é aberta pois para que um repositório tenha seus dados acessados basta que exista um tradutor capaz de acessá-lo e caso este não exista, pode ser

construído e acoplado à arquitetura. Por exemplo, no estudo de caso se fez necessário a construção de um tradutor para dados em formato Excel e seu acoplamento na arquitetura.

A tarefa de construir um tradutor para um novo tipo de dados é uma tarefa complexa quando comparada com as atividades realizadas pelos usuários da arquitetura, porém deverá ser desempenhada por profissionais de informática que apoiam a execução das atividades dos usuários. Desta forma, os usuários da arquitetura não são afetados pela complexidade de desenvolvimento de novos tradutores.

A abordagem da visão do mediador é virtual pois os dados permanecem nos repositórios locais que são acessados através de consultas. Esta abordagem permite uma interoperabilidade entre os repositórios de dados ao mesmo tempo que mantém a autonomia de cada um destes repositórios quanto a sua execução, projeto e comunicação.

A X-ARC oferece transparência de acesso aos dados através do conjunto de metadados que mantém sobre cada um dos dados disponibilizados e através dos termos de domínio que utiliza para classificá-los.

A estrutura e semântica dos dados é descrita com a ajuda dos metadados que são coletados durante o processo de registro dos dados. Desta forma, a expressividade semântica dos dados é capturada pela arquitetura através dos termos de domínio e disponibilizada aos usuários no momento que acessam ou procuram pelos dados.

Embora já existam tradutores preestabelecidos para alguns tipos de repositórios de dados, o encapsulamento do Le Select pela arquitetura permite que novos tradutores sejam construídos para atender às necessidades dos usuários. Tal fato

ocorreu no estudo de caso que necessitou do desenvolvimento de um novo tradutor para acessar dados em planilha eletrônica (Excel).

A X-ARC utiliza a linguagem XML para disponibilizar os dados espaciais e não-espaciais de forma unificada, deixando a critério do cliente que solicitou o acesso aos dados a decisão de que fontes de dados utilizar e de que forma combiná-las para utilizá-las. Portanto, independente da heterogeneidade e localização dos dados estes são apresentados de forma padronizada e unificada.

Embora possa existir um relacionamento ou ligação entre um dado espacial e um dado não-espacial retornado pela arquitetura a partir de uma consulta recebida, a os dados serão unificados pela X-ARC independente do relacionamento existir ou não, ou seja, basta que dois ou mais dados satisfaçam os critérios de consulta para que sejam unificados e retornados com único resultado pela arquitetura.

A X-ARC facilita o acesso e localização dos dados uma vez que utiliza um modelo de dados comum para representar os dados e disponibiliza uma linguagem de consulta padrão para recuperá-los. Desta forma, a utilização da arquitetura como camada de acesso a dados reduz a complexidade de desenvolvimento de aplicações que necessitam localizar e acessar dados armazenados em repositórios heterogêneos, distribuídos e muitas vezes desconhecidos. Os metadados da arquitetura auxiliam na descoberta de fontes de dados candidatas a uma necessidade do usuário (aplicação), além de fornecer a descrição, estrutura e a semântica dos dados.

Dentre os principais benefícios desta solução aplicada à Agrometeorologia destacam-se uma maior cooperação entre os pesquisadores das organizações envolvidas no trabalho, a expansão dos repositórios de dados dos usuários, a redução no custo da aquisição de dados, a otimização do tempo no tratamento e seleção dos dados e a não duplicação de dados existentes.

6.2 Trabalhos Futuros

O protótipo foi desenvolvido com o objetivo de avaliar a arquitetura proposta de publicação e verificar a realização de consultas sobre os repositórios disponibilizados pela arquitetura. Entretanto, alguns aspectos da arquitetura não foram totalmente explorados pelo protótipo e deverão ser estendidos de forma a possibilitar:

- o gerenciamento dos metadados compartilhados pela arquitetura de forma automática, permitindo que um metadado cadastrado por um nó participante seja informado automaticamente ao gerente mais próximo e assim por diante;
- implementar a cláusula "BY REGION" para o processamento de consultas de dados espaciais;
- implementar a segurança na troca dos metadados entre os gerentes e participantes utilizando técnicas de criptografia e assinatura digital, a fim de impedir a interceptação e alteração dos metadados.

Além disso, os seguintes temas, mas não apenas estes, poderão contribuir para a continuação do nosso trabalho:

- O uso de Ontologias de domínio para auxiliar no processo de publicação e localização de repositórios de dados correlacionados. A partir de uma ontologia do domínio seria possível a arquitetura propor repositórios de dados inicialmente não considerados pelos usuários da arquitetura, relacionando os repositórios a partir de termos de domínio equivalentes ou similares;

- O uso de RDF no conjunto de metadados mantidos pela arquitetura, a fim de melhorar a representação semântica dos dados disponibilizados pelos repositórios;
- O uso de termos de domínio na categorização dos atributos dos dados disponibilizados. Desta forma, durante o processo de publicação seria possível para a arquitetura identificar atributos similares e adotar uma única denominação para os atributos em repositórios diferentes e com diferente denominação;
- O uso da linguagem Geography Markup Language (GML) (GML, 2002) para representar os dados espaciais disponibilizados pela arquitetura, alcançando desta forma a mesma independência de sistemas proprietários, representação ou formato dos dados não-espaciais;
- Utilizar técnicas apropriadas para qualificar os dados disponibilizados pela arquitetura a fim de auxiliar o usuário na escolha dos dados;
- Utilizar técnicas para identificar e indicar a evolução na estrutura dos repositórios e consequentemente atualizar os tradutores que disponibilizam seus dados; e
- Acessar dados disponibilizados na Internet através de Serviços Web (Web Services), utilizando sua modularidade, independência e auto descrição para localizar serviços de acesso a repositórios e encapsular o acesso aos mesmos.

REFERÊNCIAS BIBLIOGRÁFICAS

ABITEBOUL, S., 1997, "Query Semi-Structured Data". In: *International Conference on Database Theory*, pp. 1-18, Delphi, Greece.

ARCIMS, 2001, *The ArcIMS 3 Architecture*. ESRI, J-8488.

BARU, C., GUPTA, A., LUDASCHER, B., *et al.*, 1999, "XML-Based Information Mediation with MIX". In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA, USA.

BOOCH, G., RUMBAUGH, J., JACOBSON, I., 2000, *UML Guia do Usuário*, Rio de Janeiro, Campus.

BRAY, T., 1998, "RDF and Metadata". In: http://www.xml.com/pub/a/98/06/rdf.html. Acessado em 14/10/2000.

BUSSE, S., KUTSCHE, R., LESER, U., *et al.*, 1999, *Federated Information Systems: Concepts, Terminology and Architectures*, 99-9.

CAREY, M. J., *et al.*, 1997, *Towards Heterogeneous Multimedia Information Systems: The Garlic Approach*, RJ9911.

CHAWATHE, S., GARCIA-MOLINA, H., HAMMER, J., *et al.*, 1994, "The TSIMMIS Project: Integration of Heterogeneous Information Sources". In: *Proceedings of the 10th Anniversary Meeting*, pp. 7-18, Tokyo, Japan.

DOMENIG, R., DITTRICH, K. R., 1999, "An Overview and Classification of Mediated Query Systems", *ACM SIGMOD Record*, v. 28, n. 3.

DUBLINCORE, 1999, "Dublin Core Metadata Element Set, Version 1.1: Reference Description". In: http://dublincore.org/documents/1999/07/02/dces/. Acessado em 15/08/2001.

EARTH VIEW, 2002, "Earth and Moon Viewer". In: http:// www.fourmilab.ch/earthview/vplanet.html. Acessado em 15/12/2002.

ELMAGARMID, A., RUSINKIEWICZ, M., SHETH, A., 1999, *Management of Heterogeneuos and Autonomous Database Systems*, San Francisco, CA, Morgan Kaufmann Publishers.

ELMASRI, R.,NAVATHE, S., 2000, *Fundamentals of Database Systems*, Adisson-Wesley.

FGDC, 1998, "Content Standard for Digital Geospatial Metadata". In: http://www.fgdc.gov/metadata/contstan.html. Acessado em 15/08/2001.

GAMA, E., HELM, R., JOHNSON, R., VLISSIDES, J., 2000, *Padrões de Projeto: Soluções Reutilizáveis de Software Orientado a Objetos,* Porto Alegre, Bookman.

GML, 2002, "Geography Markup Language". In: http://www.opengis.net/gml. Acessado em 14/10/2000.

GOLDMAN, R., MCHUGH, J., WIDOM, J., 1999, "From Semistructured Data to XML:Migrating the Lore Data Model and Query Language". In: *2nd International Workshop on the Web and Databases (WebDB '99)*, pp. 25-30, Philadelphia, Pennsylvania, USA.

GÜNTHER, O., LESSING, H., SWOBODA, W., 1996, "UDK: A European Environmental Data Catalogue". In: *Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modelling*, Santa Fe, New Mexico, USA.

HARDER, C., 1998, *Serving Maps on the Internet: Geographic Information on the World Wide Web*. 1, Redlands, California, USA, ESRI Press.

HTML, 1999, "HTML 4.0 Specification". In: http://www.w3.org/TR/REC-html40/. Acessado em 14/10/2000.

HURSON, A. R., BRIGHT, M. W., PAKZAD, S. H., 1994, *Multidatabase Systems: An Advanced Solution for Global Information Sharing*, Los Alamitos, CA, IEEE Computer Society Press.

KALINICHENKO, L. A., 2001, "Subject Mediation for Integrated Access to Heterogeneous Information Sources". In: *ADBIS 2001*.

KERHERVÉ, B., GERBÉ, O., 1997, "Models for Metadata or Metamodels for Data ?". In: *Proceedings of the Second IEEE Metadata Conference*, Silver Spring, Maryland.

LARRAONA, Y., MOURA, A. M. C., MATTOSO, M., 1999, "Uma Ferramenta para Gerência de Metadados em Arquiteturas Baseadas em Mediadores". In: *XIV Simpósio Brasileiro de Banco de Dados*, Florianópolis, Santa Catarina.

LESELECT, 2002, "Le Select- A Mediator System Developed in the Caravel Project". In: http://caravel.inria.fr/Fprototype_LeSelect.html. Acessado em 15/03/2000.

LIGHT, R., 1999, *Iniciando em XML*, São Paulo, Makron Books.

LUDASCHER, B., PAPAKONSTANTINOU, Y., VELIKHOV, P., 1998, "A Brief Introduction to XMAS", *Information Systems*, v. 23, n. 8.

MEDEIROS, S. P. J., 2002, *SPeCS-SISTEMA DE SUPORTE À DECISÃO ESPACIAL COLABORATIVA*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

MEDEIROS, S. P. J., STRAUCH, J. C. M., SOUZA, J. M., *et al.*, 2000, "SPECS - A SPATIAL DECISION SUPPORT COLLABORATIVE SYSTEM FOR ENVIRONMENT DESIGN". In: *Proceedings of CSCWD2000*, Hong Kong.

MEDEIROS, S. P. J., STRAUCH, J. C. M., SOUZA, J. M., *et al.*, 2001, "SPeCS - a Spatial Decision Support Collaborative System for Environment Design". In: *Procedings of SAC2001*, Las Vegas, USA.

MIX, 2002, "Mediation of Information using XML". In: http://www.db.ucsd.edu/Projects/MIX/. Acessado em 21/06/2000.

ÖZSU, M. T.,VALDURIEZ, P., 1999, *Principles of Distributed Database Systems*. Ed. 2, Upper Saddle River, New Jersey, Prentice-Hall.

PINTO, G. R. B., MEDEIROS, S. P. J., SOUZA, J. M., *et al.*, 2001, "X-Arc Spatial Data Integration in the SPeCS Collaborative Design Framework". In: *Proceedings of the Sixth International Conference on CSCW in Design*, pp. 56-60, London, Ontario, Canada.

PINTO, G. R. B., STRAUCH, J. C. M., SOUZA, J. M., *et al.*, 2002a, "Um mediador para o problema de integração de dados agrometeorológicos". In: *Anais do XII Congresso Brasileiro de Meteorologia*, pp. 1-7, Foz do Iguaçu, Brasil.

PINTO, G. R. B., STRAUCH, J. C. M., SOUZA, J. M., *et al.*, 2002b, "A Framework to Support Scientific Knowledge Management: a Case Study in Agro-meteorology". In: *Proceedings of the Seventh International Conference on CSCW in Design*, pp. 320-324, Rio de Janeiro, Brasil.

PIRES, P. F., 1997, *HIMPAR, Uma arquitetura para Interoperabilidade de Objetos Distribuídos*, Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro,RJ, Brasil.

RAMIREZ, M. R., 2001, *Processamento Distribuído de Consultas Espaciais*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

RDF, 1999, "Resource Description Framework (RDF) Model and Syntax Specification". In: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/. Acessado em 14/10/2000.

RDF, 2002, "Resource Description Framework". In: http://www.w3.org/RDF/. Acessado em 18/10/2000.

RIVEST, R. L., 1992, "RFC 1321: The MD5 Message-Digest Algorithm", *Internet Activities Board.*

RIVEST, R. L., SHAMIR, A., ADLEMAN, L. M., 1978, "A method for obtaining digital signatures and public-key cryptosystems", *Communications of the ACM*, v. 2, n. 21, pp. 120-126.

ROTH, M. T.,SCHWARZ, P., 1997, "A Wrapper Architecture for Legacy Data Sources". In: *Proceedings of VLDB 97.*

RSA LABORATORIES, 2000, "RSA Laboratories' Frequently Asked Questions About Today's Cryptography, Version 4.1". In: http://www.rsasecurity.com/rsalabs/faq. Acessado em 15/10/2002.

SAIF, 1995, "Spatial Archive and Interchange Format: Formal Definition Release 3.1". In: http://home.gdbc.gov.bc.ca/SAIF/Default.htm. Acessado em 14/08/2001.

SGML, 1986, "ISO 8879:Information processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML).", Geneva.

SHETH, A., 1998, "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics". In Goodchild, M. F., Egenhofer, M. J., Fegeas, R., and Kottman, C. A., *Interoperating Geographic Information Systems*, chapter 0, Kluwer.

SHETH, A.,LARSON, J. A., 1990, "Federated Database Systems for Managing Distributed, Heterogeneuos and Autonomous Databases", *ACM Computing Surveys*, v. 22(3), pp. 183-236.

SOUZA, J. M., 1986, *Software Tools for Conceptual Schema Integration*, Tese de Ph.D., School of Information Systems, University of East Anglia, England.

SOUZA, J. M., STRAUCH, J. C. M., PINTO, G. R. B., *et al.*, 2002, *AGROMET: Gestão do Conhecimento em Agrometeorologia*. COPPE/UFRJ, ES-582/02.

STRAUCH, J. C. M., 1998, *Integração de Bases de Dados Geográficas Heterogêneas e Distribuídas*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

TANAKA, A., VALDURIEZ, P., SIMON, E., *et al.*, 2001, "The Ecobase Project: Database and Web Technologies for Environmental Information Systems", *ACM SIGMOD Record*, v. 30(3), pp. 70-75.

TOMASIC, A., RASCHID, L., VALDURIEZ, P., 1998, "Scaling Access to Heterogeneous Data Sources with DISCO", *IEEE Transaction on Knowledge and Data Engineering*, v. 10(5), pp. 808-823.

TOMASIC, A., SIMON, E., 1997, "Improving Access to Environmental Data Using Context Information", *ACM SIGMOD Record*, v. 26, n. 1, pp. 11-15.

VIDAL, V. M. P., LÓSCIO, B. F., SALGADO, A. C., 2001, "Using Correspondence Assertions for Specifying the Semantics of XML-Based Mediators". In: *Proceedings of the International Workshop on Information Integration on the Web*, pp. 3-11, Rio de Janeiro, RJ, Brasil.

W3C, 2002, "World Wide Web Consortium". In: http://w3c.org. Acessado em 14/08/2000.

WIEDERHOLD, G., 1992, "Mediators in the Architecture of Future Information Systems", *IEEE Computer*, pp. 38-49.

XHUMARI, F., AMZAL, M., SIMON, E., 1999, "Le Select: a Middleware System for Publishing Autonomous and Heterogeneous Information Sources". In: http://caravel.inria.fr/Fprototype_LeSelect.html. Acessado em 15/06/2000.

XML, 2000, "Extensible Markup Language (XML) 1.0 (Second Edition)". In: http://www.w3.org/TR/REC-xml. Acessado em 12/04/2001.

XML, 2002, "Extensible Markup Language (XML)". In: http://www.w3.org/xml. Acessado em 08/06/2002.

APÊNDICE A – CASOS DE USO

A seguir, são apresentados casos de uso que representam as principais funcionalidades da arquitetura.

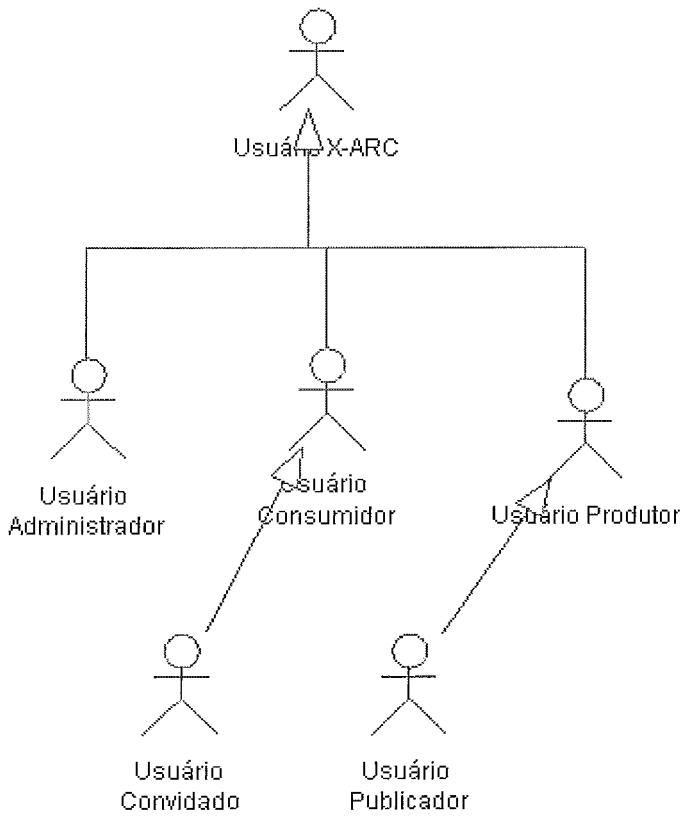


Figura A.1 – Caso de Uso Usuários

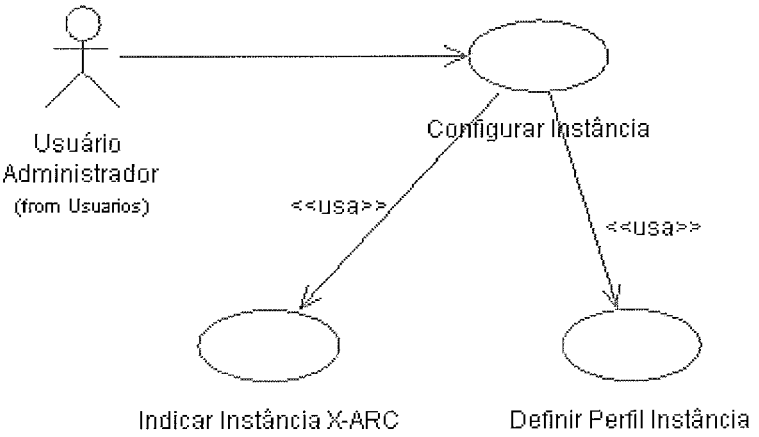


Figura A.2 – Caso de Uso Configurar Instância XARC

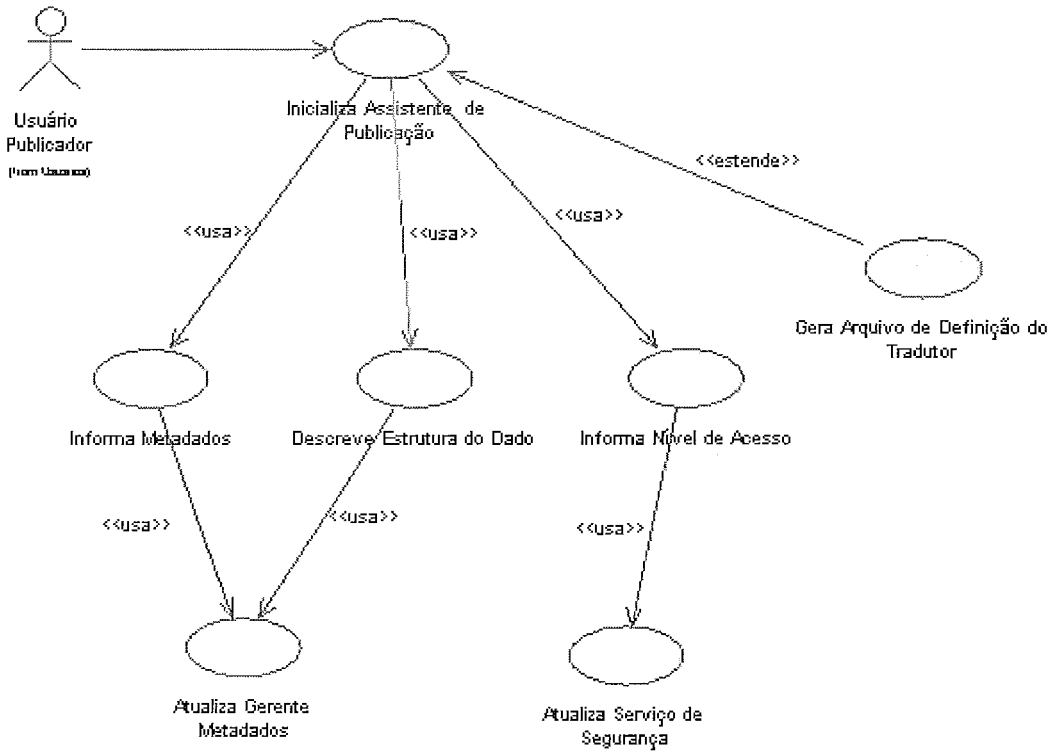


Figura A.3 – Caso de Uso Publicar Dados

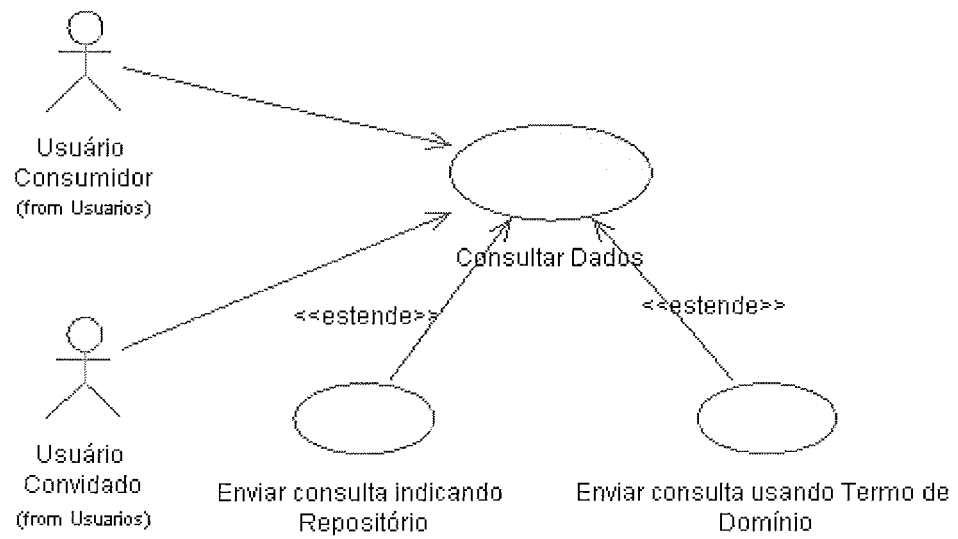


Figura A.4 – Caso de Uso Consultar Dados

APÊNDICE B – DIAGRAMA DE CLASSES DA X-ARC

A seguir, serão apresentadas as principais classes desenvolvidas durante o desenvolvimento do protótipo X-ARC. As principais classes apresentadas representam os serviços disponibilizados pela arquitetura.

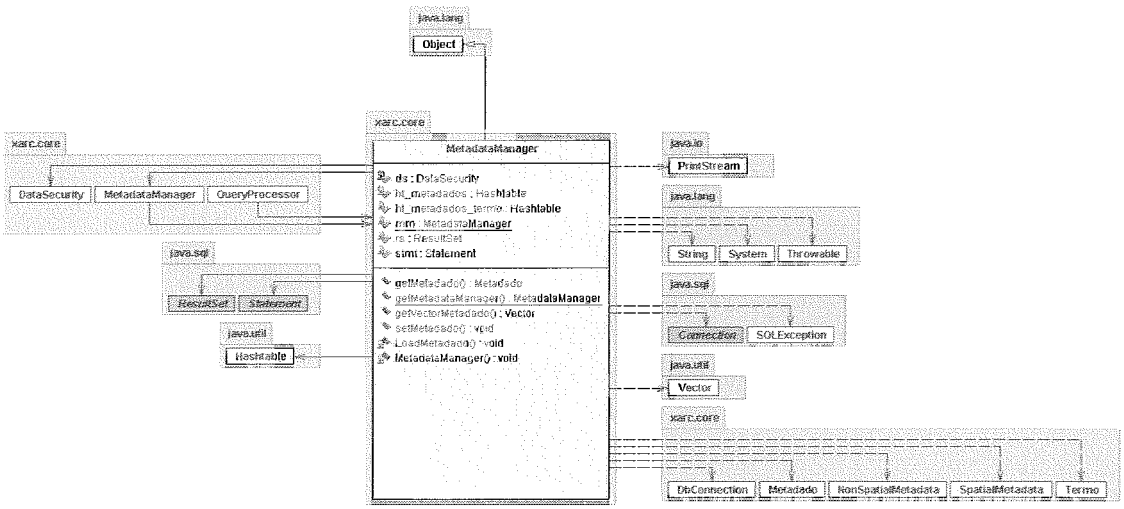


Figura B.1 – Classe MetadataManager (Gerente de Metadados)

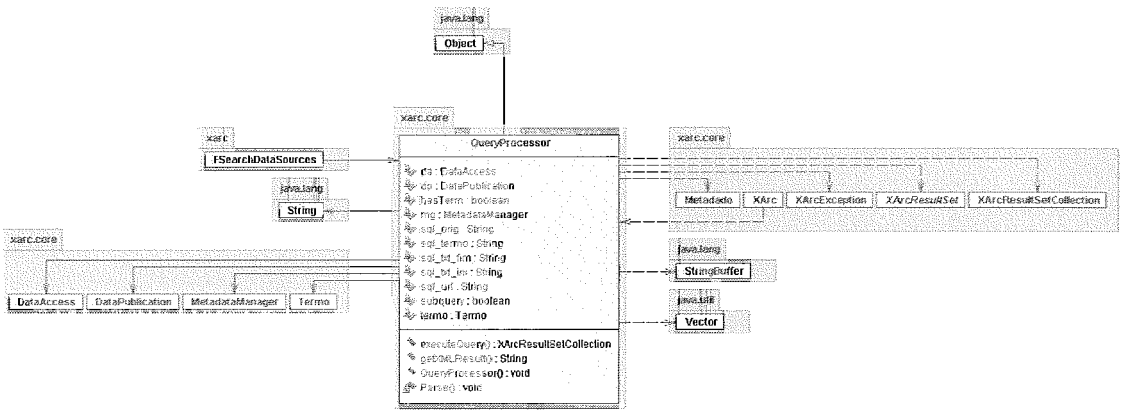


Figura B.2 – Classe QueryProcessor (Processador de Consulta)

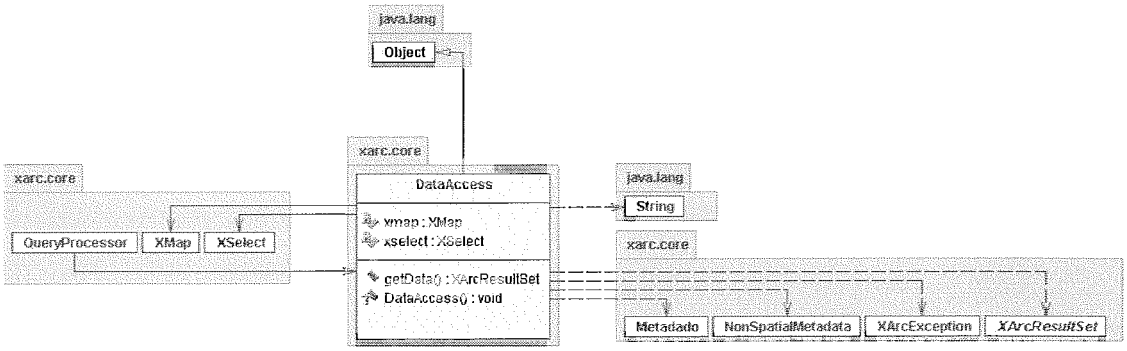


Figura B.3 – Classe DataAccess (Acesso aos Dados)

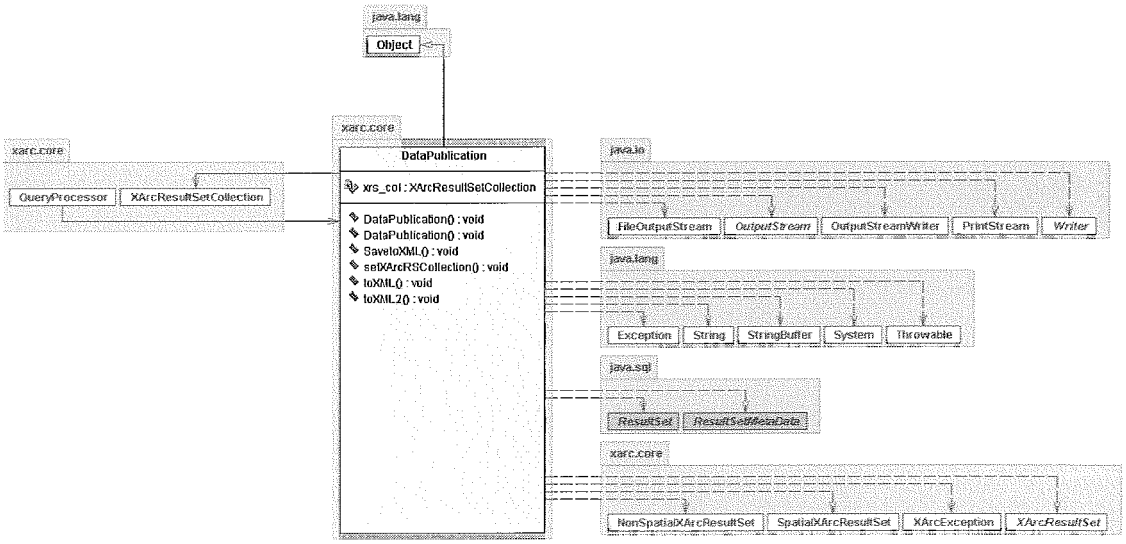


Figura B.4 – Classe DataPublication (Publicação de Dados)

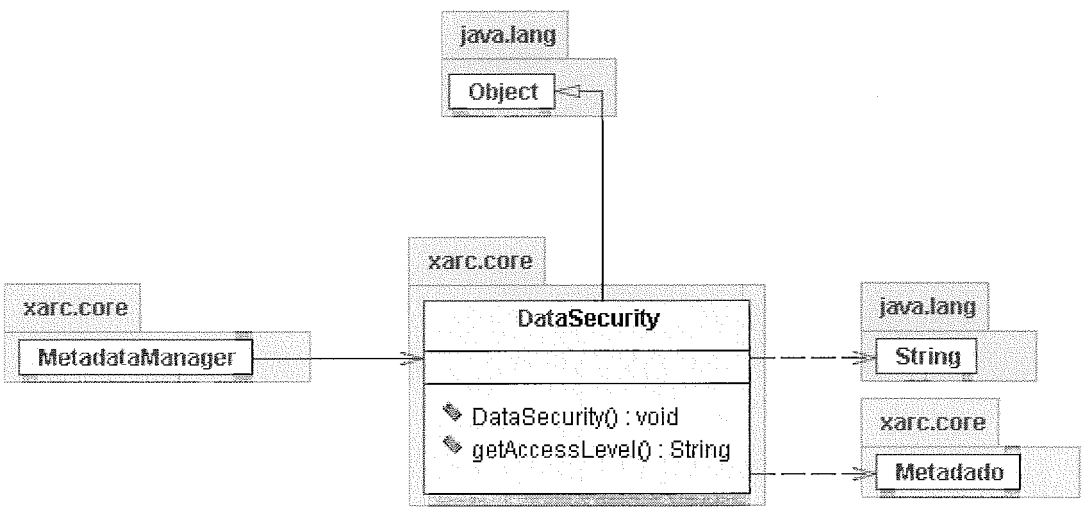


Figura B.5 – Classe DataSecurity (Segurança dos Dados)

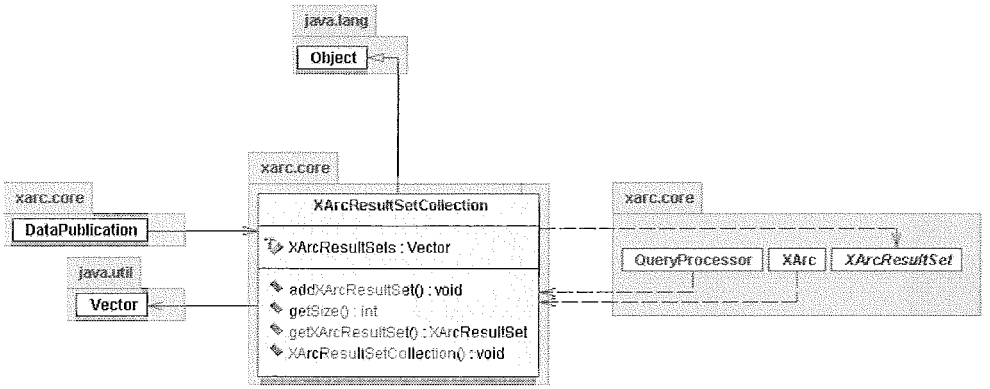


Figura B.6 – Classe XarcResultSetCollection

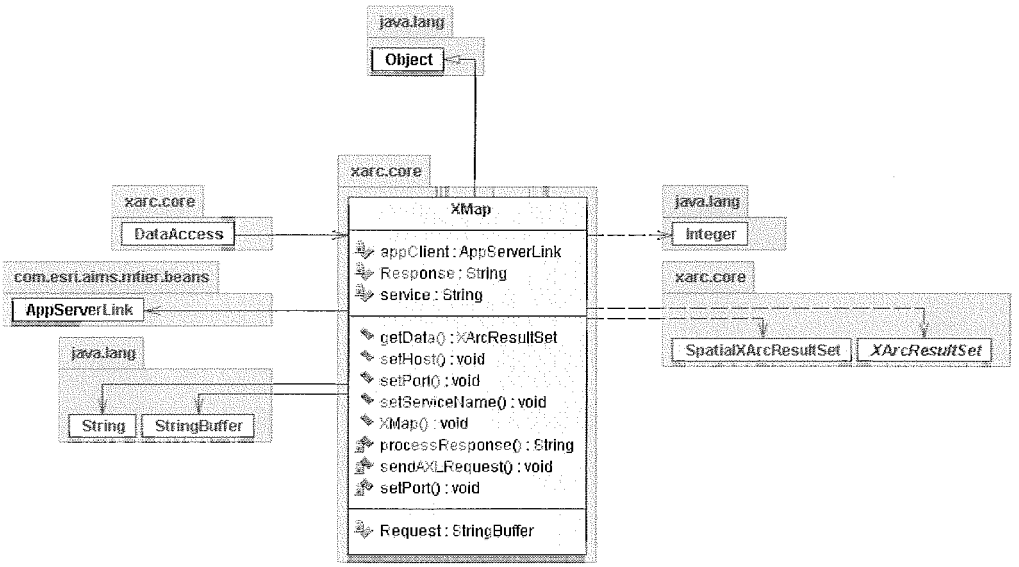


Figura B.7 – Classe XMAP

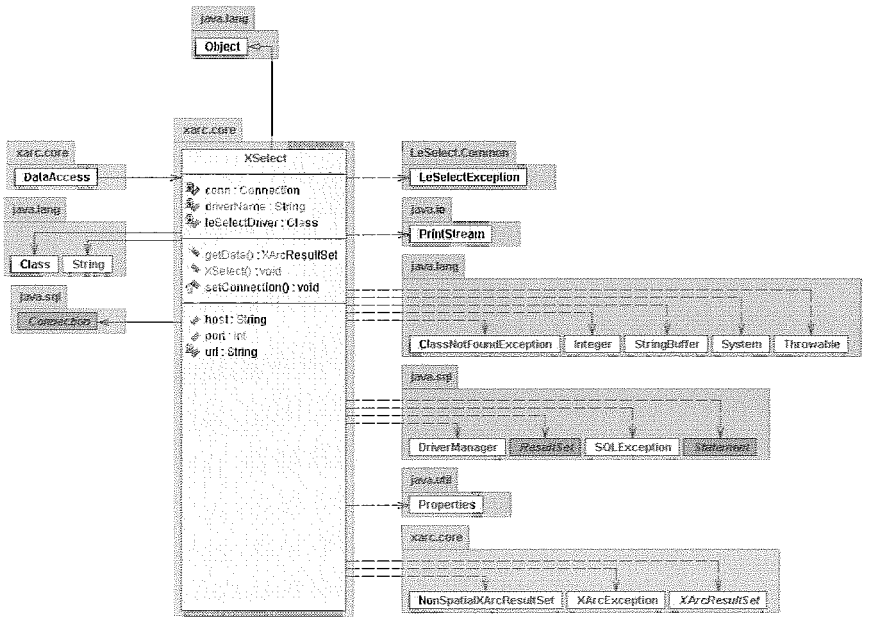


Figura B.8 – Classe XSELECT

APÊNDICE C – DIAGRAMAS DE CLASSES DO TRADUTOR EXCEL

A seguir, serão apresentadas as principais classes estendidas do Le Select que foram necessárias para o desenvolvimento de um tradutor específico para o Excel. A biblioteca "ExcelRead" foi utilizada e é representada pelo pacote **xlrd**.

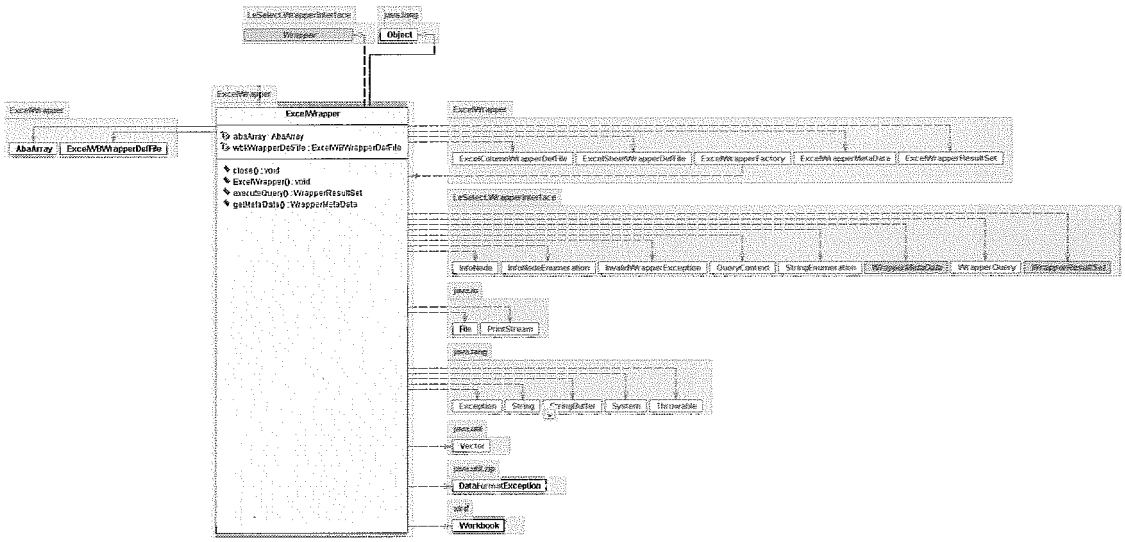


Figura C.1 – Classe ExcelWrapper

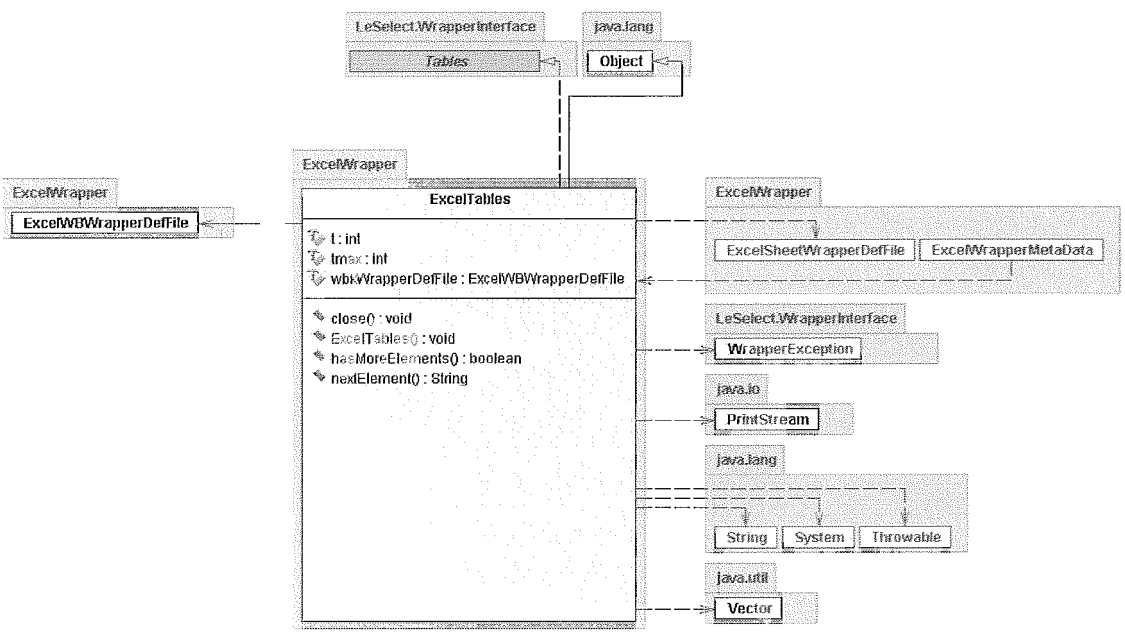


Figura C.2 – Classe ExcelTables

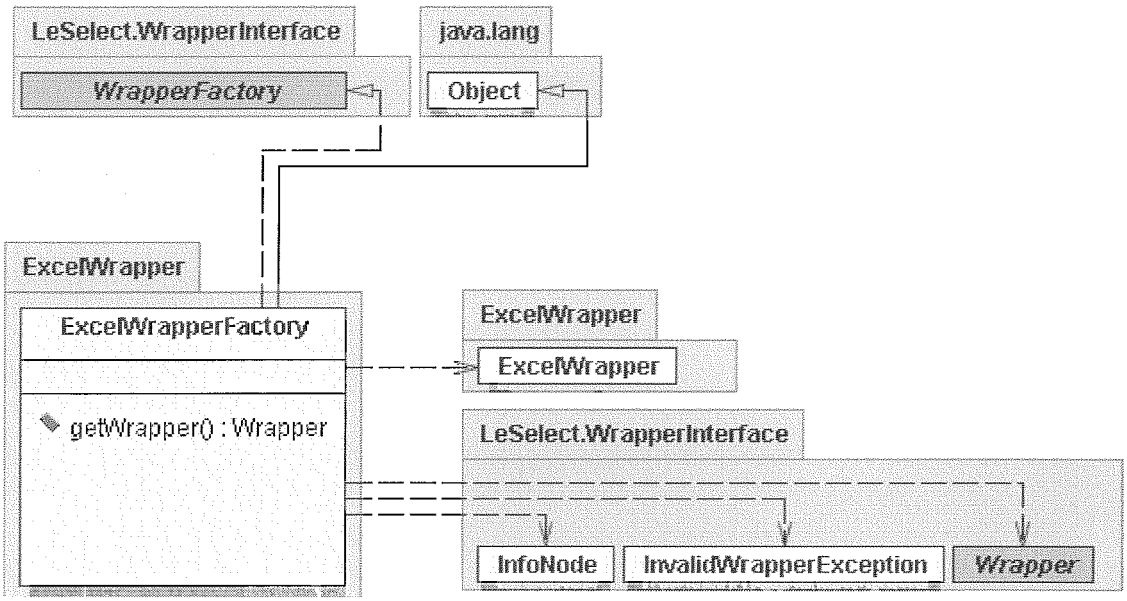


Figura C.3 – Classe ExcelWrapperFactory

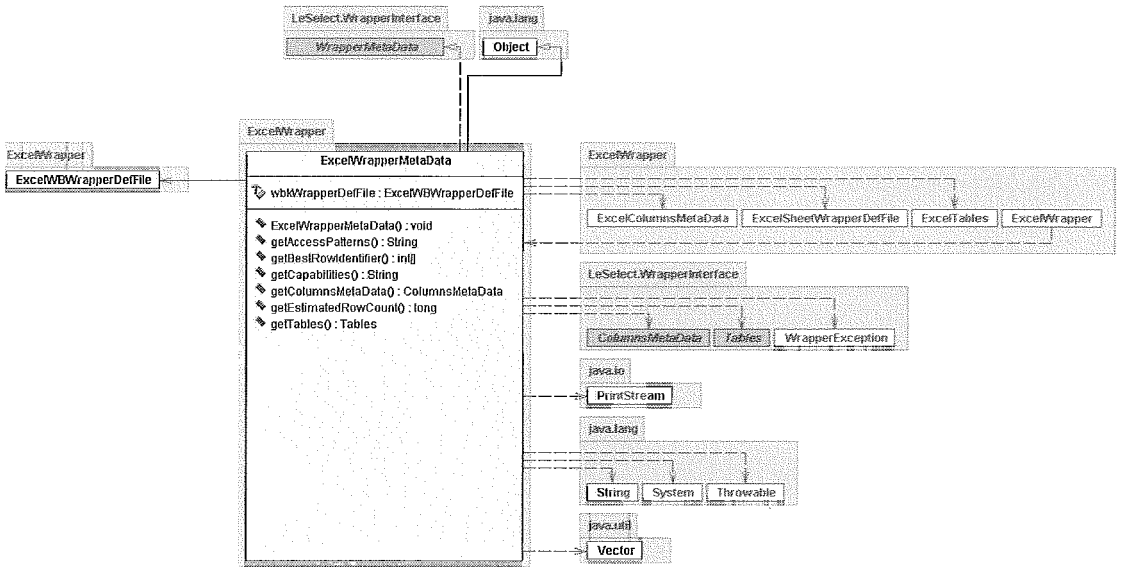


Figura C.4 – Classe ExcelWrapperMetadata

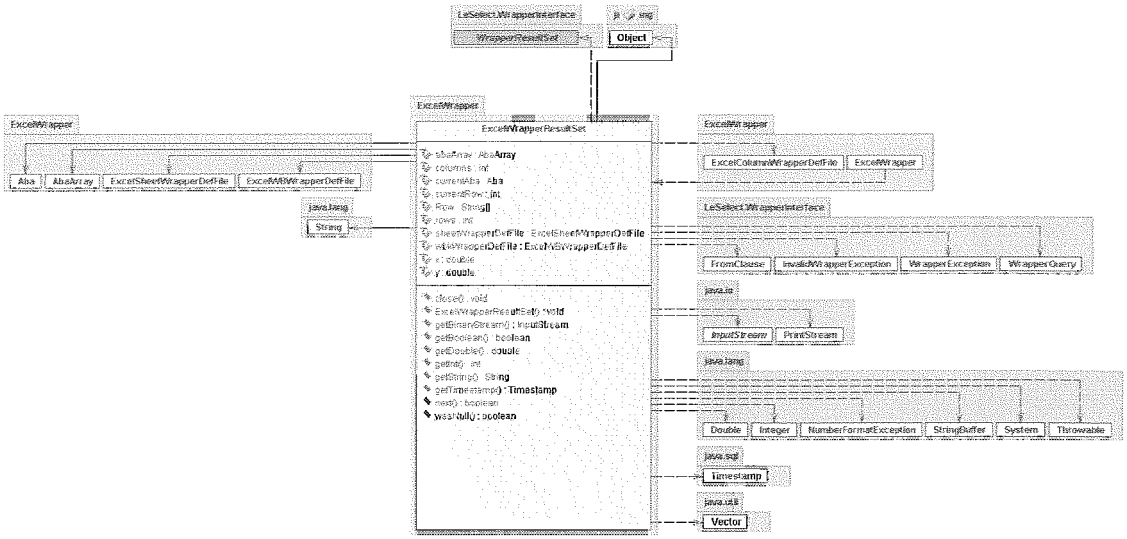


Figura C.5 – Classe ExcelWrapperResultSet

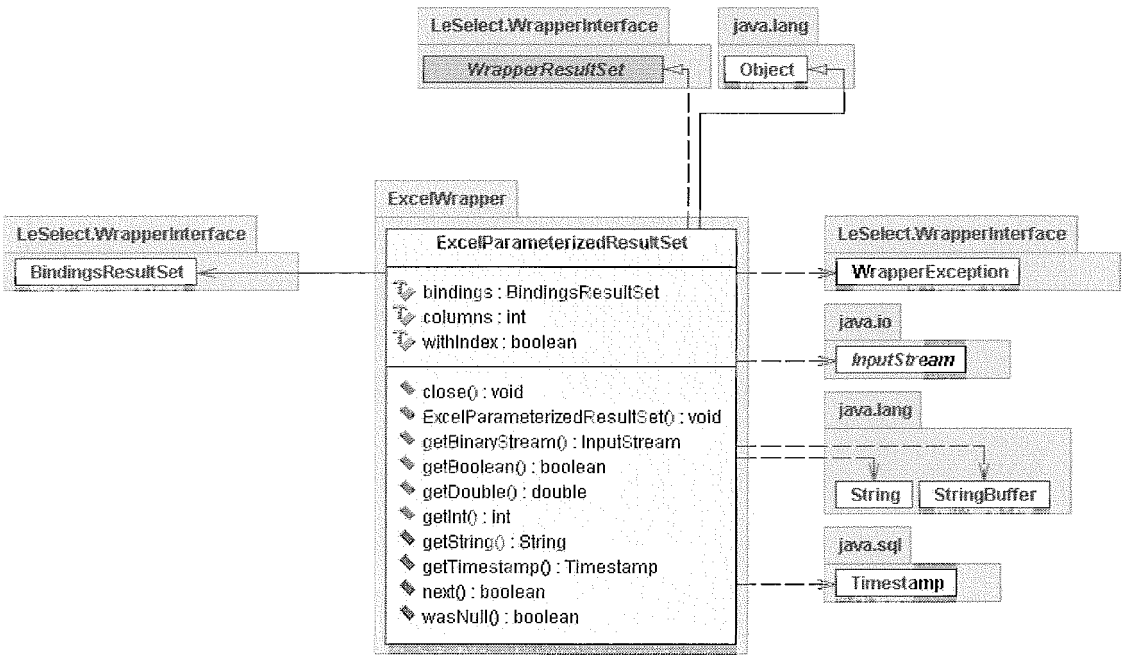


Figura C.6 – Classe ExcelParameterizedResultSet

APÊNDICE D – LISTAGEM DE CÓDIGO

Devido à extensão das linhas de código geradas para a implementação das classes do protótipo, este apêndice apresenta o código implementado para o Serviço de Processamento de Consulta.

```
package xarc.core;
import java.util.Vector;
public class QueryProcessor
{
 private String sql_txt_ini;
 private String sql_txt_fim;
 private String sql_orig;
 private String sql_termo;
 private String sql_url;
 private MetadataManager mg;
 private DataPublication dp;
 private DataAccess da;
 private boolean subquery;
 private boolean hasTerm;
 private Termo termo;

 public QueryProcessor() throws XArcException
 {
  subquery = false;
  hasTerm = false;
  sql_termo = "";
  sql_url = "";
  sql_txt_ini = "";
  sql_txt_fim = "";
  mg = MetadataManager.getMetadataManager();
  dp = new DataPublication();
  da = new DataAccess();
  termo = new Termo();
 }

 public XArcResultSetCollection executeQuery( String sql ) throws XArcException
 {
  XArcResultSetCollection xrs_col = new XArcResultSetCollection();
  Vector mds;
  Metadado md;

  // faz o parser da consulta
  Parse(sql);
  // se um termo foi utilizado na consulta, entao podem existir sub-consultas
  if (hasTerm)
  {
   // consulta no metadatamanager se existem ou nao fontes classificadas sobre o termo
   // retorna um vetor dos metadados destas fontes
   mds = mg.getVectorMetadado(termo);

   // se existem fontes entao re-escreve as subconsultas
   if (mds.size()>0 )
   {
    // para cada site é enviado uma consulta
    for(int i=0; i < mds.size();i++)
```

```
      {
        xrs_col.addXArcResultSet(da.getData(sql_txt_ini                                    +
((Metadado)mds.elementAt(i)).getUrl() + sql_txt_fim,(Metadado)mds.elementAt(i)));
      }
    }
    else //nao existem fontes para o termo
      throw new XArcException("There is no Source classified under the term \"" + termo +
"\"");
   }
   else //nenhum termo foi usado, procura a fonte para enviar a consulta
   {
    //procura o metadado no MetadataManager a partir da url da consulta
    md = mg.getMetadado(sql_url);
    if (md != null)
    {
     xrs_col.addXArcResultSet(da.getData(sql,md));
    }
    else //fonte nao existe
     throw new XArcException("There is no data in \"" + sql_url + "\"");
   }
   return xrs_col;
  }
  public String getXMLResult(String sql) throws XArcException
  {
   String file = "E:/xarc/temp/xmlresult.xml";

   XArcResultSetCollection xrs_col;
   //executa a consulta
   xrs_col = executeQuery(sql);
   dp.setXArcRSCollection(xrs_col);
   //publica o dado
   dp.SavetoXML(file);
   dp.toXML2();
   return file;
  }

  private void Parse(String consulta)
  {
   sql_orig = consulta;
   // verifica se a consulta usa Termo
   if (consulta.indexOf("<") > 0)
   {
    hasTerm = true;
    // obtem o termo de domínio utilizado
    sql_termo = consulta.substring(consulta.indexOf("<") + 1, consulta.indexOf(">"));
    termo.setValor(sql_termo);
    sql_txt_ini = consulta.substring(0,consulta.indexOf("<"));
    sql_txt_fim = consulta.substring(consulta.indexOf(">")+1, consulta.length());
   }
   // obtem o destino da consulta
   else
   {
    sql_url = consulta.substring(consulta.indexOf("//"), consulta.length()-1);
    hasTerm = false;
   }
  }
}
```