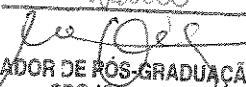


Este exemplar corresponde à redação final da
Tese/Dissertação devidamente corrigida e defendida
por: JOÃO GUILHERME DE SOUZA LIMA
e aprovada pela Banca Examinadora.
Campinas, 24 de março de 2002

COORDENADOR DE PÓS-GRADUAÇÃO
CPG-IC

635604002

**Gerenciamento de Dados Climatológicos
Heterogêneos para Aplicações em Agricultura**

João Guilherme de Souza Lima

Dissertação de Mestrado

UNICAMP
BIBLIOTECA CENTRAL
SEÇÃO CIRCULANTE

Gerenciamento de Dados Climatológicos Heterogêneos para Aplicações em Agricultura

João Guilherme de Souza Lima

Outubro de 2003

Banca Examinadora:

- Profa. Dra. Claudia Bauzer Medeiros
Instituto de Computação, UNICAMP (Orientadora)
- Prof. Dr. Rodolfo Jardim de Azevedo
Instituto de Computação, UNICAMP
- Profa. Dra. Cristina Dutra de Aguiar Ciferri
Departamento de Informática, UEM
- Prof. Dra. Anamaria Gomide
Instituto de Computação, UNICAMP (Suplente)

UNIDADE BC
1ª CHAMADA TUNICAMP
L628g
/ EX
COMBO BC/ 58116
PROC 16 - 117 - 04
C D X
PREÇO R\$ 11,00
DATA 01/06/04
Nº CPD _____

CM00197833-9

BIB ID 316774

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO IMECC DA UNICAMP**

Lima, João Guilherme de Souza
L628/18 Integração de dados climatológicos heterogêneos /João
Guilherme de Souza Lima -- Campinas, [S.P. :s.n.], 2003.

Orientadora : Claudia Bauzer Medeiros.

Dissertação (Mestrado) - Universidade Estadual de
Campinas, Instituto de Computação.

1.Banco de dados. 2.Meteorologia agrícola.
3.Sistemas de informação geográfica. I. Medeiros, Claudia
Bauzer. II. Universidade Estadual de Campinas. Instituto
de Computação. III. Título.

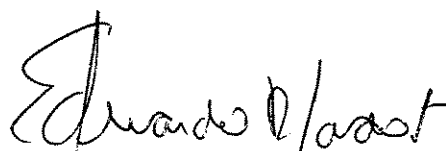
Gerenciamento de Dados Climatológicos Heterogêneos para Aplicações em Agricultura

Este exemplar corresponde à redação final da
Dissertação devidamente corrigida e defendida
por João Guilherme de Souza Lima e aprovada
pela Banca Examinadora.

Campinas, 27 de outubro de 2003.



Prof. Dra. Cláudia Bauzer Medeiros
Instituto de Computação, UNICAMP
(Orientadora)

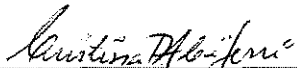


Dr. Eduardo Delgado Assad
Embrapa Informática Agropecuária
(Co-orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

TERMO DE APROVAÇÃO

Tese defendida e aprovada em 27 de outubro de 2003, pela Banca examinadora composta pelos Professores Doutores:



Profa. Dra. Cristina Dutra de Aguiar Ciferri
UEM



Prof. Dr. Rodolfo Jardim de Azevedo
IC - UNICAMP



Profa. Dra. Claudia Maria Bauzer Medeiros
IC - UNICAMP

© João Guilherme de Souza Lima, 2003.
Todos os direitos reservados.

A meus pais, a chave de tudo.

Agradecimentos

O primeiro agradecimento é para meus pais, Eduardo e Clélia. Eles possibilitaram que eu fizesse minha pós-graduação, sempre querendo o melhor pra mim. Sem minha família eu não tenho nada.

Não há como não agradecer à minha orientadora, que todo o tempo se preocupou de verdade com a qualidade do meu trabalho.

Agradeço também ao meu co-orientador, Eduardo Assad, por estar sempre disposto a discutir e validar as propostas da minha dissertação.

Quero deixar um obrigado à minha namorada, Flávia. Seus incentivos me encheram de forças para continuar.

Durante o programa tive o privilégio de participar da equipe do LIS (*Laboratory of Information Systems*). No LIS, além de ter um ambiente de trocas de idéias e discussões valiosas, pude conviver com pessoas amigas. Daniel e Fileto me ajudaram várias vezes. Nielsen, Renata, Ricardo e Gilberto também se fizeram presentes.

Agradeço à CAPES e ao projeto MCT/PRONEX SAI pelo suporte financeiro e ao Instituto de Computação pela ótima estrutura oferecida para pesquisa.

Resumo

A utilização eficiente dos recursos agrícolas requer acompanhamentos e previsões corretas dos fatores climáticos. Estes monitoramentos devem ser realizados através da análise de medidas de elementos climáticos como temperatura, precipitação e umidade do ar (os chamados *dados climatológicos*). A integração de dados de fontes diferentes possibilita melhores caracterizações dos comportamentos climáticos, fornecendo bases mais sólidas para o planejamento das atividades agrícolas e diminuindo os riscos de perdas de safras.

O objetivo deste trabalho é estudar os problemas envolvidos na integração de dados climatológicos e propor uma arquitetura que resolva estes problemas. A arquitetura proposta trata questões de homogeneização de dados, avaliação de qualidade e disponibilização de dados. Metadados descritivos facilitam o acesso aos dados, e resultados de avaliações de qualidade provêm mais suporte às tomadas de decisão baseadas nas informações obtidas do sistema. O trabalho foi validado através da implementação de parte da arquitetura no contexto de um projeto de monitoramento agroclimatológico desenvolvido na Empresa Brasileira de Pesquisa em Agropecuária (EMBRAPA).

As principais contribuições desta dissertação são: (1) estudo dos problemas inerentes à integração de dados climatológicos; (2) estudo e levantamento das questões relativas à avaliação de qualidade de dados; (3) proposta de uma arquitetura de integração de dados climatológicos e (4) validação da arquitetura através de implementação parcial para um caso real.

Abstract

The efficient use of agricultural resources requires monitoring and correct forecasts of climatic factors. These monitorings must be made through analysis of measures of climatic elements (known as *climatological data*) such as temperature, precipitation and air humidity. The integration of data from various sources enables better characterization of climatic behavior, helping to plan agricultural activities and decreasing the chances of crop losses.

The objective of this dissertation is to study the problems related to climatological data integration and to specify an architecture that solves these problems. The proposed architecture deals with data homogenization, quality evaluation and data availability. Descriptive metadata facilitate data access, and results of quality evaluations provide additional support to the decision process. This work was validated through partial implementation of the architecture for a climatological monitoring project developed in the Brazilian Agricultural Research Corporation (EMBRAPA).

The main contributions of this research are: (1) study of problems inherent to climatological data integration; (2) study and survey of questions associated with climatological data quality evaluation; (3) specification of a climatological data integration architecture and (4) validation of the architecture through partial implementation for a real case study.

Sumário

Agradecimentos	xi
Resumo	xiii
Abstract	xv
1 Introdução	1
2 Revisão Bibliográfica	5
2.1 Sistemas de Informação Geográficos e bancos de dados espaço-temporais	5
2.2 Integração de bancos de dados heterogêneos	7
2.2.1 Problemas na integração de dados geográficos	7
2.3 Intercâmbio de dados	9
2.3.1 Sistemas de Bancos de Dados Federados	10
2.3.2 Metadados	11
2.3.3 Linguagens para intercâmbio de dados na Web	13
2.3.4 Ontologias	17
2.3.5 Representação de ontologias	19
2.4 Qualidade de dados	22
2.5 Séries históricas e dados climatológicos	26
2.5.1 Dados pluviométricos	27
2.6 Resumo	31
3 Problemas e Requisitos do Sistema	33
3.1 Questões a serem consideradas	33
3.1.1 Heterogeneidade	33
3.1.2 Avaliação de qualidade	35
3.1.3 Disponibilização dos dados	38
3.2 Requisitos do sistema	39
3.2.1 Visão geral	39

3.2.2	Requisitos de pré-processamento	40
3.2.3	Requisitos de consultas	42
3.3	Resumo	43
4	Arquitetura Proposta	47
4.1	Arquitetura do sistema	47
4.2	Banco de Dados	49
4.3	Banco de Metadados	51
4.3.1	Relacionamento entre dados e metadados	52
4.3.2	Padrões utilizados	53
4.3.3	Metadados de Identificação	54
4.3.4	Metadados de Cobertura Espacial	54
4.3.5	Metadados de Cobertura Temporal	56
4.3.6	Metadados de Unidades	57
4.3.7	Metadados Administrativos	57
4.4	Módulo de Integração de Dados	58
4.4.1	Tradutores de dados	59
4.4.2	Migrador de dados	60
4.5	Módulo de Avaliação de Qualidade	61
4.5.1	Indicadores de qualidade utilizados	62
4.5.2	Avaliação não-dependente e dependente da tarefa	63
4.6	Módulo de Processamento de Consultas	64
4.6.1	Funções básicas	65
4.6.2	Resultados de consultas	65
4.7	Resumo	68
5	Aspectos de implementação	69
5.1	Banco de Dados	69
5.2	Módulo de Integração	74
5.3	Módulo de Avaliação de Qualidade	75
5.3.1	Consistência de séries históricas	76
5.3.2	Avaliação não-dependente da tarefa	79
5.3.3	Visões	79
5.4	Dificuldades da implementação	81
5.5	Resumo	82
6	Conclusões e extensões	83
6.1	Contribuições	83
6.2	Extensões	84

6.2.1	Extensões da arquitetura	84
6.2.2	Extensões de implementação	85
A	Esquema do banco de metadados	87
B	Esquema XML para os dados homogeneizados	89
	Bibliografia	98

Lista de Tabelas

2.1	Dimensões de qualidade de dados [56].	24
4.1	Unidades de medidas e precisões do esquema global.	49
4.2	Granularidades de tempo para cada tipo de medida no esquema global. . .	50
4.3	Metadados de Identificação.	55
4.4	Metadados de Cobertura Espacial.	56
4.5	Metadados de Cobertura Temporal.	56
4.6	Metadados de Unidades.	57
4.7	Metadados Administrativos.	58

Lista de Figuras

2.1	Uma definição de esquema em XML Schema.	14
2.2	Um exemplo simples de um documento RDF.	16
2.3	Seção <i>ontology-container</i> de uma ontologia OIL.	20
2.4	Seção <i>ontology-definitions</i> de uma ontologia OIL.	21
2.5	Exemplo de aplicação do método dupla-massa.	29
3.1	Exemplo de estatísticas a serem exibidas para anos.	44
3.2	Exemplo de estatísticas a serem exibidas para meses.	45
4.1	Diagrama da arquitetura proposta.	48
4.2	Diagrama entidade-relacionamento do esquema de dados global.	52
4.3	Relacionamento entre dados e metadados.	53
4.4	Módulo de Integração de Dados.	59
4.5	Exemplos de resultados da avaliação não-dependente da tarefa.	64
4.6	Interação entre os módulos de Avaliação de Qualidade e de Processamento de Consultas.	66
5.1	Diagrama entidade-relacionamento.	70
5.2	Esquema resultante do diagrama entidade-relacionamento.	73
5.3	Exemplo de arquivo de dados no formato original da instituição.	74
5.4	Exemplo de arquivo de saída dos tradutores.	75
5.5	Tela de estimativa de medidas faltantes.	78
5.6	Visões implementadas.	80
A.1	Esquema do banco de metadados.	88

Capítulo 1

Introdução

A agricultura tem uma importância significativa na economia brasileira. Estima-se que 30% do Produto Interno Bruto do Brasil advenha das atividades da agropecuária e do agrobusiness, ou seja, dos negócios da agricultura [3]. Produtos como soja, algodão e café são essenciais, inclusive, à exportação e à manutenção da balança comercial brasileira.

Apesar de lucrativa, a atividade agrícola também envolve riscos e gera prejuízos quando enfrenta o caráter aleatório dos fatores climáticos. Tome-se como exemplo a perda de 34% da safra 92/93 de arroz no estado da Bahia. Em casos como a safrinha de milho, os dispêndios já alcançaram 120 milhões de reais anuais.

As perdas de safras podem ser amenizadas com a coleta e análise de dados climatológicos. A organização e o processamento eficiente destes dados possibilitam extrações de informações que, por sua vez, permitem a agricultores e cooperativas um melhor planejamento de suas atividades. Eventos climáticos como veranicos, chuvas extremas e geadas arrasam safras se não previstos. Por outro lado, a previsão correta de tais eventos possibilita a tomada de medidas preventivas capazes de diminuir ou mesmo anular os prejuízos.

Existem no Brasil diversas instituições que coletam dados climatológicos. O potencial de extração de informação não é totalmente aproveitado, uma vez que cada organização utiliza apenas seus próprios dados. A integração de dados climatológicos de fontes diferentes possibilita melhores caracterizações de regiões, e, conseqüentemente, previsões e identificação de cultivos propícios mais corretas.

Porém os dados climatológicos no Brasil apresentam uma série de problemas causados pela sua heterogeneidade. Como as instituições que coletam os dados trabalham independentemente entre si, tendo cada uma definido suas próprias formas de coleta e armazenamento, não existe no Brasil um padrão de modelagem e disponibilização de dados que seja largamente difundido e aceito.

No contexto de dados climatológicos, destacam-se as *séries históricas de medidas plu-*

viométricas. Tais séries permitem estimar probabilidades de chuvas e ocorrências de períodos chuvosos e não-chuvosos para regiões determinadas. Com o estudo dos padrões de chuva de uma região, pode-se definir locais que tenham condições pluviométricas semelhantes e agrupá-los convenientemente em sub-regiões. Esta classificação, aliada a uma caracterização do tipo de solo, permite determinar que tipo de cultivo é mais propício, quando é adequado plantá-lo e os riscos associados.

Os objetivos desta dissertação são estudar os problemas envolvidos na integração de dados climatológicos heterogêneos, e especificar a arquitetura de um sistema de integração de tais dados. Esta arquitetura deve abordar questões relativas à avaliação de qualidade dos dados.

A solução proposta foi validada através da implementação parcial de um sistema de integração de dados pluviométricos, que, como mencionado, correspondem a um tipo específico de dados climatológicos. A implementação foi realizada no contexto de um projeto real desenvolvido na Empresa Brasileira de Pesquisa em Agropecuária (EMBRAPA). Órgãos como o Ministério da Agricultura, a Agência Nacional de Águas e o Instituto Nacional de Pesquisas Espaciais alimentam o sistema com séries históricas pluviométricas, algumas com 100 anos de dados.

No projeto da EMBRAPA, o sistema de integração de dados pluviométricos faz parte de um sistema maior, chamado **Agritempo**. O objetivo do Agritempo é funcionar como um sistema de monitoramento agrometeorológico para todo o Brasil. Este sistema irá disponibilizar, na Web, diversos produtos informativos (dados, mapas e informações em geral) aos agentes do agronegócio, incluindo informações referentes a chuva, disponibilidade de água no solo, previsões climáticas e condições para manejo do solo. Agricultores podem utilizar tais produtos como base para tomadas de decisão em diferentes fases do cultivo de diversas culturas, auxiliando, por exemplo, a programação do plantio, colheita e secagem de produtos, a aplicação de defensivos agrícolas e de adubos foliares, o controle preventivo ou combate direto a geadas, entre outros.

As principais contribuições desta dissertação são:

- estudo e levantamento dos problemas inerentes à integração de dados climatológicos heterogêneos;
- levantamento e discussão das questões envolvidas em avaliação de qualidade de dados;
- proposta de um conjunto de metadados para descrever dados climatológicos;
- proposta de uma arquitetura de integração de dados climatológicos;
- proposta de uma metodologia de avaliação de qualidade de dados; e

- implementação parcial da arquitetura e discussão das dificuldades encontradas na implementação.

Esta discussão deu origem a uma publicação na conferência GEOInfo [25].

O restante desta dissertação está organizado da seguinte forma. O capítulo 2 define os conceitos necessários ao entendimento do texto. O capítulo 3 discute as questões envolvidas no projeto de um sistema de integração de dados climatológicos, e apresenta alguns dos requisitos de usuário levantados na EMBRAPA. A seguir, o capítulo 4 descreve a arquitetura proposta para um sistema de integração de dados climatológicos. O capítulo 5 apresenta os aspectos da implementação realizada para validar a arquitetura proposta. Finalmente, o capítulo 6 apresenta as conclusões da dissertação, e propõe algumas extensões para o trabalho.

Capítulo 2

Revisão Bibliográfica

A integração de dados a que este trabalho se propõe exige estudo em diferentes aspectos: sistemas geográficos, bancos de dados espaço-temporais, integração de dados e intercâmbio de conhecimento. Este capítulo apresenta os conceitos básicos iniciais para a dissertação, e discute algumas das abordagens utilizadas hoje para enfrentar problemas das áreas citadas.

A seção 2.1 define e apresenta algumas características de Sistemas de Informação Geográficos e de bancos de dados espaço-temporais. Em seguida, a seção 2.2 discute problemas relacionados à integração de bancos de dados heterogêneos, dedicando atenção especial a dados geográficos. A seção 2.3 apresenta três abordagens para intercâmbio e troca de dados entre fontes diferentes. A seção 2.4 introduz questões relacionadas à avaliação de qualidade de dados. A seguir, a seção 2.5 caracteriza dados climatológicos, foco deste trabalho. Por fim, a seção 2.6 resume o capítulo.

2.1 Sistemas de Informação Geográficos e bancos de dados espaço-temporais

Sistemas de Informação Geográficos (SIGs) têm tido um papel importante nas últimas décadas. Eles são utilizados em atividades de planejamento urbano, agricultura, silvicultura, agropecuária e outros campos. SIGs têm se difundido em todo o mundo, à medida em que o universo de usuários e as oportunidades de treinamento na tecnologia se expandem e à medida em que novas aplicações deste tipo de sistema são descobertas.

Há várias definições de SIGs na literatura. Elas variam de acordo com a ênfase desejada, que pode estar voltada aos requisitos funcionais, aplicações suportadas, tipos de dados tratados, etc. SIGs também podem ser vistos como sistemas compostos por vários subsistemas integrados voltados à geração de mapas e extração de informações sobre da-

dos geográficos. Este trabalho utiliza a definição de [27], onde SIGs são definidos como sistemas de informação que tratam de dados de alguma maneira geograficamente referenciados, isto é, dados com localização espacial definida referencialmente à Terra.

Nesta dissertação, considera-se que dados georeferenciados são dados *espaço-temporais*. Tais dados possuem atributos *espaciais* (que determinam suas características geográficas), atributos *temporais* e atributos *convencionais*. Os atributos convencionais podem variar sem haver modificação dos atributos espaciais, e vice-versa.

No contexto deste trabalho, os dados espaço-temporais manipulados consistem em medidas climatológicas diárias ou horárias. Neste caso, o atributo espacial é a localização geográfica do local onde a medida foi coletada. Um dos atributos convencionais é a medida climatológica propriamente dita (temperatura ou umidade do ar, por exemplo), e o atributo temporal é a data e, se for o caso, a hora da medida.

Um banco de dados que lida com dados espaço-temporais é chamado de *banco de dados espaço-temporal*. O esquema de um banco de dados espaço-temporal possui a definição dos atributos convencionais, espaciais e temporais. A noção de tempo é obtida armazenando-se um novo registro cada vez que o valor de um atributo convencional ou espacial é alterado.

Há a opção de armazenar, a cada modificação ocorrida, um registro completo, ou apenas o atributo cujo valor sofreu alteração. Esta opção determina o *nível de marcação de tempo* do banco de dados. Além disto, a marca de tempo associada ao registro pode ser de dois tipos: *tempo válido* ou *tempo de transação*. O tempo válido corresponde ao tempo no qual os atributos espaciais e convencionais são válidos. O tempo de transação corresponde ao momento em que os dados foram armazenados no banco de dados. Em aplicações geográficas geralmente utiliza-se somente o tempo válido.

Outro fator a considerar é a *granularidade temporal*. Trata-se da unidade de tempo escolhida como base para a marcação da variação dos dados. Esta escolha depende das aplicações que se pretende desenvolver usando o sistema. Determinados valores podem ser medidos com periodicidades diferentes (por exemplo, dia, mês, ano ou estação do ano), o que pode provocar alguns problemas no armazenamento de dados temporais. Por exemplo, a amostragem pode ser irregular, pode haver intervalos de tempo sem medidas.

Com isto surge o problema de se obter valores de atributos relativos a instantes de tempo para os quais não há registro no banco de dados (a chamada *interpolação temporal*). Pode-se utilizar funções que tentam deduzir este valor a partir de outros instantes de tempo para os quais há registro. O desenvolvimento destas funções não é trivial e pode apresentar problemas e dificuldades diversas [27]. Em aplicações geográficas, muitas vezes se considera que o valor de um atributo é constante desde o tempo em que ele foi considerado válido até a próxima marca de tempo armazenada. A seção 2.5.1 apresenta alguns dos métodos propostos na literatura para inferir medidas pluviométricas faltantes.

2.2 Integração de bancos de dados heterogêneos

Ambientes heterogêneos de bancos de dados são definidos na literatura de diversas formas. De maneira geral, o termo *bancos de dados heterogêneos* é utilizado para referenciar um conjunto de bancos de dados com diferentes SGBDs, modelos de dados, esquemas ou semântica dos dados. Estas diferenças se referem ao nível de banco de dados, mas podem ocorrer também no hardware (plataformas), nos protocolos de comunicação, ou nas aplicações, que ajudam a definir a semântica dos dados. A interpretação dos dados pode mudar, mesmo que eles se refiram à mesma realidade.

Conflitos de modelos de dados ocorrem, por exemplo, quando os bancos de dados utilizam diferentes modelos de dados - por exemplo, quando um banco de dados utiliza o modelo relacional e outro o modelo de orientação a objetos. Neste caso, a resolução consiste em converter cada esquema para um modelo de dados comum, resolvendo também diferenças de restrições de integridade existentes entre os modelos.

Conflitos de esquema ocorrem devido às alternativas diferentes providas por um modelo de dados para desenvolver esquemas para uma mesma realidade. Por exemplo, o que é modelado como um atributo em um esquema pode ser modelado como uma relação em outro esquema para o mesmo domínio de aplicação. Outro exemplo é o uso de nomes diferentes para designar elementos que se referem ao mesmo tipo de objeto no mundo real. A identificação de correspondências pode requerer conhecimento adicional sobre o esquema, o modelo de dados empregado e o domínio de aplicação. Outro problema está relacionado às restrições dos esquemas de cada banco de dados componente (cada banco de dados sendo integrado), e que são parte da semântica.

A resolução dos conflitos de semântica se baseia na determinação e padronização do significado de conceitos, termos e estruturas encontrados nas origens de dados. Como já constatado em [66], o nível de heterogeneidade mais difícil de superar para atingir integração é o nível de semântica. Ao contrário de outros níveis que se resumem a problemas técnicos, para se chegar a um consenso de significado dos dados é necessária a interação entre pessoas familiarizadas com os dados sendo integrados. A seção 2.3.4 apresenta uma abordagem que vem sendo proposta para lidar com heterogeneidade no nível semântico: a estruturação do conhecimento na forma de *ontologias*.

2.2.1 Problemas na integração de dados geográficos

Um dos problemas enfrentados em SIGs é o grande volume de dados heterogêneos gerados, com pouco compartilhamento de terminologia, semântica e padrões de armazenamento. Como resultado, não é simples fazer com que sistemas geográficos compartilhem dados, nem com que usuários treinados em um sistema se tornem aptos a operar outros sistemas. Devido às peculiaridades dos dados geográficos, a integração de bancos de dados

geográficos se mostra mais difícil do que a integração de bancos de dados relacionais convencionais.

A grande diversidade de aplicações geográficas torna impossível a definição de um padrão para lidar com seus vários aspectos [52]. Há sistemas formais para definir localizações, mas não para definir objetos mais complexos. Por exemplo, o conceito de latitude e longitude é um padrão universal, mas não há uma linguagem universal utilizada para definir todos os objetos e eventos da superfície terrestre. Significados comuns existem apenas dentro de um mesmo domínio de estudos. O item mais difícil de se alcançar em interoperabilidade de SIGs é conseguir compartilhar a semântica de dados armazenados em sistemas diferentes. Como citado, a comunidade de pesquisa em sistemas geográficos vem explorando o uso de ontologias para tal integração.

O trabalho de Fonseca [36] apresenta um estudo sobre integração de informações geográficas, onde é introduzido o conceito de *papéis*. Este conceito se refere ao fato de que fenômenos geográficos podem se modificar ao longo do tempo e também serem vistos como coisas diferentes por diferentes grupos de pessoas. Assim, um objeto pode ter distintos papéis em um mesmo instante ou ao longo de sua vida, dependendo do ponto de vista de um grupo de usuários. Desta forma, papéis podem ser utilizados como uma “ponte” entre os vários níveis de detalhe em uma estrutura de conceitos.

Não há um consenso para se chegar a um conjunto de requisitos funcionais, em termos de operadores espaciais e relações a serem utilizados sobre dados espaciais. Isto ocorre devido à grande variedade de domínios de aplicação e variedade de usuários, que possuem diferentes necessidades e expectativas quanto ao que um sistema geográfico deve fornecer [47]. Operadores e relações espaciais dependem de vários fatores, como escala, tempo, ponto de vista e precisão da sua especificação. Como não há um padrão para lidar com estes fatores, cada sistema os trata de maneira diferente.

Um tipo específico de heterogeneidade que pode ocorrer em SIGs se refere ao nível de marcação de tempo utilizado pelos sistemas sendo integrados. Assim, uma perturbação na integração de dados acontece quando eles utilizam níveis de marcação de tempo diferentes, ou seja, quando um sistema armazena um registro completo para cada modificação ocorrida e outro armazena apenas o atributo que sofreu alteração. Conjuntos de dados também podem variar quanto à granularidade temporal.

Outro tipo particular de heterogeneidade está relacionada à granularidade espacial de cada sistema sendo integrado. Por exemplo, a comparação entre dados de um sistema que armazena dados na escala 1:10.000 e outro que utiliza escala 1:100.000 é muitas vezes impossível, sendo preciso recorrer à chamada *generalização cartográfica* para tentar iniciar o processo de solução [68].

Em alguns países, como os Estados Unidos, há iniciativas para integrar dados geográficos de órgãos diferentes. Um exemplo é o US Global Research Program, uma co-

laboração estabelecida entre agências norte-americanas e instituições de pesquisa [16]. O objetivo é disponibilizar a cientistas ambientais um serviço, via Internet, que sirva como um catálogo de informações ambientais relativas a mudanças globais.

Uma outra iniciativa, esta mundial, para permitir interoperabilidade em sistemas geográficos, é o *Open GIS Consortium* (OGC) [18]. O OGC é um consórcio internacional de companhias, agências de governo e universidades, com o objetivo de desenvolver um conjunto público de especificações, requisitos e padrões para permitir interoperabilidade de SIGs, e assim acelerar o sucesso comercial de geoprocessamento distribuído. As especificações do OGC definem interfaces comuns que visam a integração de dados geoespaciais e recursos de geoprocessamento através da infra-estrutura de informação global.

No contexto de dados climatológicos, destaca-se também a iniciativa da *World Meteorological Organization* (WMO) [5] para estabelecer padrões de intercâmbio de dados e de produtos hidrológicos para todo o mundo. Além do estabelecimento de padrões, a WMO também busca incentivar a livre e irrestrita troca internacional de dados hidrológicos, e auxilia países a recuperar registros antigos de dados hidrológicos.

2.3 Intercâmbio de dados

Quando se tem diversas fontes de dados e deseja-se trabalhar com todos os dados como se estes constituíssem uma base de dados única, pode-se forçar os dados a seguirem um padrão. Ou seja, esta abordagem define um modelo comum, para o qual os dados heterogêneos são transformados e a seguir integrados. O resultado é uma base de dados homogênea.

Uma alternativa a este procedimento é permitir que cada conjunto de dados permaneça na sua forma original. Neste caso, cada fonte de dados gerencia seus próprios dados e os disponibiliza para as outras fontes. Uma solução deste tipo ocorre, por exemplo, quando se opta por um *Sistema de Bancos de Dados Federados* (SBDF), onde os conjuntos de dados podem até mesmo ser gerenciados por SGBDs diferentes.

Uma outra situação onde dados são mantidos no seu formato original ocorre quando empresas e organizações disponibilizam seus dados, gratuitamente ou não, na World Wide Web. Neste caso o potencial de intercâmbio de informação é extremamente alto.

Porém, o conhecimento disponível na World Wide Web não está sendo totalmente aproveitado. Sistemas não são capazes de utilizar informações de outros sistemas correlatos devido, entre outras razões, às limitações impostas pela maneira como os dados são modelados e armazenados. Além disto, o sucesso da Internet acabou criando dificuldades para a sua própria utilização: sistemas de gerenciamento de dados e de documentos, ao lidar com enormes quantidades de informação, não conseguem oferecer ao usuário mecanismos realmente eficientes de consulta. A informação desejada, apesar de disponível,

não é atingida pelo sistema de busca. Faz-se necessário uma melhor indexação do conhecimento.

O armazenamento de descrições dos dados juntamente com os próprios dados pode ajudar a resolver este problema, e com isto aprimorar os resultados de buscas na Internet. Tais dados descritivos são chamados de *metadados* [42]. Eles possibilitam que sejam fornecidas informações sobre os dados desejados, facilitando sua localização.

Um campo emergente de pesquisa é o desenvolvimento da *semantic Web* [34, 10], que visa aumentar o compartilhamento das informações disponíveis na Internet. Neste campo procura-se alcançar uma troca não só coerente mas também automatizada de *conhecimento*, e não simplesmente de dados, entre os sistemas conectados pela Internet. Aliada à falta de uniformidade dos dados, há o problema de semântica e sua interpretação. Isto exige aumentar a consciência semântica que os sistemas têm sobre os dados armazenados.

Uma tentativa de solucionar este problema é a utilização de *ontologias* [65]: representações de um domínio de conhecimento de maneira sistemática, consensual e, no caso da computação, também inteligível para computadores. A pesquisa em ontologias dedica um cuidado especial ao modo como a semântica dos dados é definida. Conceitos são definidos e organizados de maneira que seja possível inferir as relações entre eles.

Esta seção apresenta cada uma destas três técnicas de intercâmbio de dados e de conhecimento: Sistemas de Bancos de Dados Federados, utilização de metadados e modelagem de conceitos na forma de ontologias. Também são apresentadas as principais linguagens utilizadas hoje para manipulação de metadados e de ontologias.

2.3.1 Sistemas de Bancos de Dados Federados

Um SBDF pode ser definido como “um conjunto de bancos de dados independentes, possivelmente heterogêneos, que compartilham dados entre si” [61]. A principal característica de SBDFs é a autonomia dos bancos de dados que os compõem. Em um SBDF, a integração não chega ao nível de transações, se resumindo ao nível de compartilhamento dos dados já armazenados, em consultas globais [7].

Existem três principais abordagens para implementar SBDFs. Elas não são baseadas na integração física dos dados, mas sim na integração lógica, mantendo, portanto, a autonomia de cada banco de dados participante da federação. Elas consistem em utilizar mediadores, tradutores/adaptadores e visões.

Um *mediador* é um software responsável por permitir a comunicação e interoperabilidade entre diferentes SGBDs. As consultas ao banco de dados heterogêneo são submetidas ao mediador, que as divide em subconsultas a serem enviadas a cada SGBD componente. O mediador gera estas subconsultas na linguagem de consulta de cada SGBD componente. O resultado de cada subconsulta é traduzido para a linguagem de consulta do

SGBD componente que gerou a consulta, e o resultado pode ser visto pelo usuário.

Tradutores/adaptadores convertem os modelos de dados de cada sistema componente para um modelo de dados comum, global. Com base nesta tradução, as consultas globais são convertidas em consultas específicas das fontes de informação envolvidas na consulta. Tradutores/adaptadores freqüentemente são confundidos com mediadores, mas estas ferramentas possuem diferenças. Tradutores/adaptadores são utilizados para construir o sistema unificado de dados, não fazendo a mediação entre os bancos de dados participantes; uma vez utilizados, podem ser dispensados. Já o mediador está sempre presente na comunicação entre componentes, e pode utilizar tradutores/adaptadores para resolver uma parte específica da conversão de dados.

Visões são um outro mecanismo que auxilia a integração de sistemas componentes de um SGBD. Em SGBDs relacionais, visões são encaradas como relações virtuais não normalizadas, definidas em função de relações preexistentes e obtidas como resultado do processamento de uma consulta [15]. Os dados da visão, isto é, os dados que representam o resultado desta consulta, podem ser armazenados fisicamente, constituindo uma *visão materializada*, ou podem ser obtidos somente a cada utilização da visão, sendo neste caso uma *visão virtual*. Em SGBDs, visões fornecem ao usuário uma interface a partir da qual ele pode acessar os dados heterogêneos de maneira transparente, sem se preocupar de quais sistemas componentes estão sendo obtidos os dados.

Um ou mais dos sistemas a serem integrados pode ser um *sistema legado*. Tal sistema geralmente consiste em um sistema de grande porte, antigo, autônomo, e que resiste a evoluções e modificações. Quando um dos sistemas a integrar é um sistema legado, a análise exige maior atenção. As opções possíveis para se proceder à integração dependem da organização e abertura de tal sistema. Se o sistema legado for fechado e não permitir um processo de engenharia reversa, pode-se ter que reconstruir o sistema legado para tentar migrar seus dados [62]. Há também a possibilidade de migrar os dados do sistema legado para um *sistema destino* que possua as mesmas funcionalidades do sistema legado, mas que suporte modificações de forma mais fácil e organizada.

Há outras questões também importantes na configuração de sistemas integrados, como *distribuição*, técnicas de *recuperação de informação* e *segurança*. Elas não são contempladas neste trabalho.

2.3.2 Metadados

Em diversas situações da vida cotidiana utilizamos informação para encontrar informação. Por exemplo, em grandes livrarias utilizamos sistemas de busca para descobrir a localização de um livro na loja, fornecendo dados como autor, título ou editora. Estes dados associados a outros dados a fim de descrevê-los são o que chamamos de **metadados**. No

exemplo anterior, *autor* é um metadado associado ao livro que procuramos. Metadados são, então, definidos como *dados sobre dados* [42]. Eles podem ser utilizados para descrever o conteúdo, qualidade, condição ou outras características dos dados, auxiliando inclusive buscas na Internet.

No momento de compartilhar e localizar dados na Internet através de metadados, surge um problema quando organizações utilizam terminologias diferentes para descrever os mesmos tipos de entidades. A principal iniciativa para tentar amenizar este problema é o *Dublin Core Metadata Initiative* (DC) [2]. Esta iniciativa consiste em um fórum aberto com o objetivo de desenvolver uma conjunto de metadados que suporte uma larga faixa de propósitos e modelos de negócios. O padrão de metadados Dublin Core [1] é disponibilizado gratuitamente em seu site.

É reconhecido que o padrão Dublin Core não atenderá as necessidades de todos os usuários, especialmente quando os propósitos forem muito especializados. Em algumas aplicações pode ser necessário refinar ou qualificar o significado de alguns metadados. Nest sentido, alguns qualificadores podem ser utilizados para refinar o significado de metadados do conjunto básico Dublin Core. Por exemplo, o metadado DC.Date pode ser refinado para DC.Date.created. Comunidades locais de usuários podem utilizar o Dublin Core como o ponto de partida e a seguir desenvolver suas próprias extensões.

De fato, isto já ocorre em algumas áreas. Um dentre vários exemplos é o *EdNA Metadata Standard* [8], um padrão de metadados baseado no Dublin Core criado especificamente para atender às necessidades do domínio de educação e treinamento na Austrália. Outro exemplo é uma iniciativa da Food and Agriculture Organization of the United Nations para estabelecimento de padrões de metadados para a área da agricultura [53, 37].

Existe ainda um padrão de metadados voltado a dados geoespaciais. Trata-se do padrão *Content Standards for Digital Geospatial Metadata* [30], do Federal Geographic Data Committee (FGDC). Este padrão estabelece terminologias para metadados de dados geoespaciais, e permite [59]: (1) determinar a *disponibilidade* de conjuntos de dados geoespaciais; (2) determinar a *adequabilidade* de conjuntos de dados geoespaciais; (3) determinar as *formas de acesso* a conjuntos de dados geoespaciais; e (4) *transferir* conjuntos de dados geoespaciais. Relacionado ao contexto de dados climatológicos, destaca-se também o padrão de metadados proposto pela WMO [54, 5] para intercâmbio de dados meteorológicos.

Estes esforços em estabelecer padrões demonstram os benefícios que podem ser obtidos ao utilizar metadados para intercâmbio de conhecimento e informações. Porém, além de estabelecer padrões de terminologias para metadados, é necessário definir meios de armazenar e manipular dados e metadados na Web, visando interoperabilidade. A seção a seguir apresenta duas linguagens criadas com estes propósitos.

2.3.3 Linguagens para intercâmbio de dados na Web

Esta seção apresenta duas linguagens, *XML* e *RDF*, criadas para o uso de dados semi-estruturados na Web.

XML

Em fevereiro de 1998 o World Wide Web Consortium (W3C) divulgou a primeira versão da linguagem *Extensible Markup Language* (XML) [19], voltada para a construção de documentos Web. O objetivo era permitir a criação de documentos que contivessem indicações a respeito de seu conteúdo. Em outras palavras, XML passou a permitir a elaboração de documentos contendo não só dados mas também metadados. XML é hoje um padrão na construção de documentos Web.

Em XML o conteúdo de documentos, que inclui dados e metadados, é organizado através de etiquetas. XML provê facilidades para a definição destas etiquetas e das relações estruturais entre elas, porém não especifica semântica nem um conjunto pré-definido de etiquetas. A semântica de um documento XML é definida pelas aplicações que o processam.

Esta liberdade de criação de elementos e atributos pode criar problemas de vocabulário, como colisões e dificuldade de reconhecimento. Isto pode ocorrer se documentos de um mesmo domínio de aplicação utilizam termos diferentes ou se documentos de domínios diferentes utilizam os mesmos termos.

Para amenizar este problema foi criado o mecanismo *XML namespaces*. Um XML namespace é uma coleção de nomes, identificada por uma URI (Uniform Resource Identifier) [17], para ser utilizada em documentos XML como tipos de elementos e nomes de atributos. Desta maneira documentos semelhantes podem utilizar nomes universais.

Não é objetivo deste texto discutir aspectos da sintaxe XML. Referências a conferências, livros e artigos sobre XML podem ser encontradas em [19].

XML Schema

XML Schema Definition Language, ou, simplesmente, XML Schema, é uma linguagem para definir a estrutura e conteúdo de documentos XML. A XML Schema foi aprovada como uma recomendação do W3C em maio de 2001, e provê meios para a definição de *esquemas* para documentos XML. Ela permite definir, entre outros, domínios de valores, cardinalidades e regras para etiquetas aninhadas.

XML Schema também se aproveita do mecanismo de namespaces, já que definições em XML Schema são elas próprias documentos XML. Desta maneira, aplicações desenvolvidas para XML, como ferramentas de validação, podem ser aplicadas a definições de esquema em XML Schema.

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="livro">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="titulo" type="xs:string" />
        <xs:element name="autor" type="xs:string" />
        <xs:element name="personagem"
          minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="nome" type="xs:string" />
              <xs:element name="amigo-de" type="xs:string"
                minOccurs="0" maxOccurs="unbounded" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
      <xs:attribute name="isbn" type="xs:string" />
    </xs:complexType>
  </xs:element>
</xs:schema>

```

Figura 2.1: Uma definição de esquema em XML Schema.

A Figura 2.1 exibe um exemplo simples de uma definição de esquema em XML Schema. Vejamos alguns de seus aspectos básicos. O elemento de nome `livro` é definido como sendo de tipo complexo, uma vez que ele não terá apenas um valor mas sim subelementos. Seus subelementos são definidos utilizando a restrição `sequence`, que indica que a sua ordem é relevante. Alguns dos subelementos, como `autor` e `titulo`, são definidos como do tipo simples, isto é, sem subelementos, e o seu tipo de dados (`string`, `date`, etc.) também é definido. O subelemento `personagem` é definido como do tipo complexo, e sua cardinalidade é definida de maneira que um livro possa ter nenhum ou ilimitados personagens.

Além de utilizar definições de esquema em XML Schema, também pode-se definir a estrutura de documentos XML utilizando *XML Document Type Definitions* (DTDs) [19]. DTDs definem os elementos e blocos de construções válidos para um documento XML. Grupos independentes podem utilizar DTDs para intercâmbio de dados. Por exemplo, a FGDC, que criou o padrão de metadados geospaciais apresentado na seção 2.3.2, propõe também um DTD para tal conjunto de metadados.

Uma explicação mais detalhada de XML Schema pode ser obtida em [28].

RDF

RDF, *Resource Description Framework*, é uma linguagem desenvolvida pelo W3C visando possibilitar uma utilização eficiente de metadados na Internet [50]. Ela é uma especificação de como *recursos* devem ser descritos em XML. Conseqüentemente, documentos RDF são documentos XML. RDF permite descrever e intercambiar dados associados a metadados, e é voltada para diversas áreas de aplicação, como descrição de recursos, classificação de conteúdo, comércio eletrônico e serviços colaborativos.

Bray [11] lista as quatro regras baseadas nas quais RDF foi definido:

1. Um **Recurso** é qualquer coisa que possa ser referenciado por uma URI, o que inclui páginas Web e elementos individuais de documentos XML.
2. Uma **Propriedade** é um Recurso que possui um nome e que pode ser usado como uma característica de um Recurso. Por exemplo, *autor* ou *titulo*. Na maioria dos casos o que realmente importa é o seu nome, mas uma Propriedade precisa ser um Recurso para que ela possa ter suas próprias propriedades.
3. Uma **Declaração** consiste na combinação de um Recurso, uma Propriedade e um valor. Assim, uma declaração associa um valor a uma Propriedade de um Recurso. Por exemplo, o valor da propriedade *autor* do recurso <http://www.books.com/Hamlet.html> pode ser “Shakespeare”. Um valor também pode ser um outro Recurso.
4. Há um método direto para expressar Propriedades em XML.

RDF utiliza a facilidade de XML namespaces. Um XML namespace permite a um documento RDF definir o escopo e identificar unicamente um conjunto de propriedades. Este conjunto de propriedades, chamado de *esquema*, pode ser acessado na URI correspondente ao namespace.

A Figura 2.2 ilustra as características básicas de um documento RDF. A primeira linha indica dois namespaces, RDF e DC, sendo RDF o namespace padrão. As propriedades utilizadas no documento são definidas em ao menos um destes namespaces. A seção entre as etiquetas <Description> define quatro propriedades para o recurso <http://musicas.br/Azul.html>, apontado pela URI especificada no atributo *about*. Em outras palavras, este exemplo define o valor de quatro *metadados* para uma página HTML. Vale notar que os metadados com o prefixo DC pertencem ao padrão Dublin Core.

A linguagem RDF possui diversos outros meios de descrição de metadados que não serão cobertos neste texto. A intenção aqui é apenas mostrar a aparência e a sintaxe básica de um documento RDF. Documentação mais completa de RDF pode ser obtida em [50].

```

<RDF xmlns = "http://w3.org/TR/1999/PR-rdf-syntax-19990105#"
  xmlns:DC = "http://purl.org/DC#">
  <Description about = "http://musicas.br/Azul.html">
    <DC:Title> Azul da cor do mar </DC:Title>
    <DC:Creator> Tim Maia </DC:Creator>
    <DC>Date> 1990-10-01 </DC>Date>
    <DC:Subject> Música </DC:Subject>
  </Description>
</RDF>

```

Figura 2.2: Um exemplo simples de um documento RDF.

Poder-se-ia perguntar por que simplesmente não utilizar XML ao invés de RDF. XML permite criar nomes de etiquetas conforme a semântica da aplicação. Etiquetas XML podem conter tanto dados em texto quanto outras etiquetas. Além disso, XML possui correspondentes a Propriedades e Declarações de RDF, como em ``. Assim, XML realmente pode ser utilizada para intercâmbio de metadados na Web. Porém, ela falha em um dos objetivos alcançados por RDF: escalabilidade. Isto ocorre por duas razões.

Primeiro, porque a ordem na qual elementos aparecem em um documento XML é significativa. No contexto de metadados isto não deveria ter importância. Além disto, manter a ordem correta de milhões de dados é caro e difícil, na prática.

Segundo, devido à liberdade de forma existente em XML, que permite construções como:

```

<Descricao> O valor desta propriedade contém texto misturado com propriedades
filhas como sua temperatura (<Temp>26</Temp>) e longitude (<Long>98</Long>).
[&Disclaimer;]</Descricao>

```

A representação em memória deste tipo de documento pode exigir estruturas de dados complexas que misturam árvores, grafos e cadeias de caracteres, o que não ocorre com RDF.

RDF Schema

RDF permite definir declarações sobre Recursos, usando Propriedades e valores. Entretanto, é interessante que as diversas comunidades tenham a capacidade de descrever tipos específicos de recursos, e de utilizar propriedades também específicas para descrever estes recursos. Por exemplo, uma companhia de equipamentos de camping pode querer definir

uma *classe* Barraca, e utilizar propriedades como `modelo`, `pesoEmKg`, e `tamanhoEmbalada` para descrevê-la.

RDF por si só não provê um vocabulário para especificar estes termos. Tais classes e propriedades devem ser descritas em um *vocabulário* RDF a ser criado para cada contexto de aplicação. A linguagem *RDF Schema* [12] foi criada com o objetivo de descrever estes vocabulários RDF.

Porém, RDF Schema não provê vocabulários orientados à aplicação, mas sim os mecanismos necessários para *especificar* classes e propriedades como parte de um vocabulário, e para indicar quais classes e propriedades espera-se que sejam usadas em conjunto. Desta maneira, RDF Schema provê um sistema de tipos para RDF. Este sistema de tipos lembra em alguns aspectos sistemas de tipos de linguagens de programação orientadas a objeto [12]. Por exemplo, Recursos podem ser definidos como instâncias de uma ou mais classes, e as classes podem ser definidas de maneira hierárquica, com classes inferiores herdando propriedades de classes superiores. Porém, em outros aspectos classes e propriedades RDF são bem diferentes do conceito de tipos de linguagens de programação. Não é objetivo deste texto se aprofundar nestas questões; uma descrição detalhada de RDF Schema, juntamente com exemplos de sua utilização, pode ser obtida em [12]. Uma introdução básica a XML, RDF, XML Schema e RDF Schema pode ser obtida em [40].

O W3C provê na Internet um vocabulário RDF de recursos e propriedades, juntamente com seus significados, que podem ser utilizados para descrever classes e propriedades específicas de usuário. Este vocabulário RDF Schema é definido em um namespace identificado pela URI <http://www.w3.org/2000/01/rdf-schema#>.

Vale ressaltar que, ao utilizar metadados para descrever dados da maneira como é feito em XML e RDF, não se inclui semântica nos dados. RDF e XML não auxiliam o computador a entender a semântica do que ele está processando. Isto pode ser alcançado com o uso de ontologias.

2.3.4 Ontologias

Acredita-se que a representação formal do conhecimento tenha começado na Índia, no primeiro milênio a.C., com o estudo da gramática sânscrita [60, 57]. Esta disciplina está ligada a trabalhos realizados em filosofia na Grécia antiga, principalmente por Aristóteles (384-322 a.C.). Como mencionado, ontologias consistem em representações de um domínio de conhecimento de maneira sistemática e inteligível para computadores. Na computação, ontologias foram utilizadas inicialmente no campo da Inteligência Artificial, visando facilitar o compartilhamento de conhecimento e o reuso [32].

Qualquer que seja a maneira escolhida para construir uma ontologia, deve-se armazenar, juntamente com as definições de conceitos, as relações existentes entre os mesmos -

seus relacionamentos semânticos. Assim, geralmente uma ontologia é formada por conceitos, relacionamentos entre os conceitos e axiomas.

Geralmente os conceitos do domínio sendo modelado são representados por *classes*. A ontologia toma a forma de uma árvore hierárquica, na qual cada classe herda as características da classe imediatamente superior, podendo também herdar características de mais de uma classe. As classes irmãs podem ou não ser mutuamente exclusivas, dependendo do caso. Tem-se como resultado uma taxonomia de noções na qual o significado de um nó é dado por todas as propriedades, similaridades e diferenças encontradas no caminho do conceito raiz (o mais genérico) até o nó [9].

Além de herança, outros relacionamentos também podem ser definidos na ontologia, entre quaisquer pares de classes. Por exemplo, o relacionamento *instrumento* entre os conceitos *empresaAerea* e *turismo*. O armazenamento deste relacionamento permitiria inferir que empresas aéreas tornam possível a realização da atividade turismo.

Axiomas definem regras sobre os relacionamentos. Por exemplo, um axioma pode definir se um relacionamento entre dois conceitos é simétrico ou não. Um outro tipo de axioma é a composição de relacionamentos. O relacionamento *AvoDe*, por exemplo, poderia ser definido pela composição de dois relacionamentos *PaiDe*. Staab e Maedche [63] propõem uma abordagem de modelagem de axiomas que visa especificá-los independentemente da linguagem a ser utilizada.

Ao mesmo tempo em que é necessário definir os relacionamentos entre os conceitos de maneira clara e sem ambiguidade para permitir seu correto processamento, também é importante que os usuários possam visualizar e entender a ontologia. Por isso algumas abordagens suportam a modelagem de ontologias em várias camadas. A camada superior geralmente corresponde ao que um ser humano consegue entender, e utiliza linguagem natural. Desta maneira o usuário pode varrer a ontologia, seja para consultá-la, modificá-la ou validá-la manualmente. Já a camada inferior deve ser definida mais formalmente, para fazer a interface com o computador. As camadas intermediárias formam mapeamentos entre as outras camadas. Este foi o tipo de abordagem utilizado na especificação da linguagem OIL [38], descrita mais adiante, na seção 2.3.5.

Algumas abordagens não se contentam com apenas fornecer meios para modelagem e armazenamento de ontologias, mas também tentam automatizar pelo menos parcialmente estes processos. É o caso das ferramentas que realizam *aprendizado automatizado de conceitos*. Geralmente, estas ferramentas analisam documentos e Web sites relacionados ao domínio da aplicação e extraem uma terminologia para o domínio. Então, *processadores de linguagem natural* filtram as informações obtidas e detectam relações taxonômicas entre os conceitos aprendidos.

Este tipo de análise não é trivial, uma vez que é essencial haver interpretação semântica para capturar relacionamentos entre os conceitos aprendidos. Um destes relacionamentos,

por exemplo, é a generalização *tipo-de*. Ela poderia ocorrer, por exemplo, entre os conceitos *metrô* e *transporte público*. Um exemplo de utilização de técnicas de aprendizado automatizado é o ambiente desenvolvido com base na ferramenta OntoLearn [51], que suporta não só aprendizado mas também validação automatizada de conceitos.

No processo de criação de uma ontologia para um domínio de conhecimento pode-se tentar aproveitar partes de outras ontologias já existentes. Neste sentido, um interessante sistema de referência online e gratuito é o WordNet [31]. Ele consiste em um banco de dados léxico da língua inglesa, e sua construção foi inspirada por teorias psicolinguísticas da memória léxica humana. Substantivos, verbos, adjetivos e advérbios em inglês são organizados em conjuntos de sinônimos, cada um representando um conceito léxico. No WordNet, as palavras podem ter mais de um significado, e cada significado de cada palavra é definido de duas maneiras: com um conjunto de termos chamado de *synset* e com uma definição textual chamada de *gloss*. O *synset* contém termos relacionados ao significado em questão. Por exemplo, em um dos significados do verbo *ship*, o *synset* consiste nos termos *transport*, *send* e *ship*. Desta maneira conceitos correlatos podem ser relacionados através de termos em comum nos seus *synsets*.

2.3.5 Representação de ontologias

A representação de ontologias em sistemas de computador exige linguagens capazes de expressar conhecimento de maneira clara e não-ambígua. Desde os anos 90 têm havido algumas propostas com este objetivo.

Em bancos de dados, a solução correspondente é o uso de modelos e esquemas para definir a estrutura e parte da semântica dos dados. Assim, pode ser natural comparar ontologias com definições de esquemas. Fensel [32] lista as seguintes diferenças entre estes dois tipos de modelagens de dados:

- uma linguagem para definição de ontologias é sintática e semanticamente mais rica que abordagens comuns para bancos de dados;
- a informação que é descrita por uma ontologia consiste em textos de linguagem natural semi-estruturados e não de informação tabular;
- uma ontologia precisa ser uma terminologia consensual e compartilhada, uma vez que ela deve ser utilizada para intercâmbio e compartilhamento de informação; e
- uma ontologia provê uma teoria de domínio e não a estrutura de um repositório de dados.

```

ontology-container
  title "Animais africanos"
  creator "Ian Horrocks"
  subject "Animais, comida e vegetarianos"
  description "Um exemplo didático de ontologia que descreve animais africanos."
  description.release "1.01"
  publishser "Ian Horrocks"
  type "ontology"
  format "pseudo-xml"
  identifier "http://www.exemplos.org/oil/exemplo.onto"
  source "http://www.africa.com/animais.html"
  language "OIL"
  language "pt-br"
  relation.hasPart "http://www.ontosRus.com/animais/selva.onto"

```

Figura 2.3: Seção *ontology-container* de uma ontologia OIL.

Estas comparações podem levar a entender que especificações de esquemas de bancos de dados são imprecisas, o que não é verdade. A seguir são apresentadas três linguagens utilizadas hoje para modelagem de ontologias: OIL, DAML+OIL e OWL.

Ontology Inference Layer (OIL) [38] é uma linguagem criada para representar semântica de uma maneira acessível por máquinas, modelando domínios de conhecimento na forma de ontologias. Desenvolvida para ser compatível com padrões do W3C, incluindo XML e RDF, OIL explora as primitivas de modelagem de RDF Schema. Desta maneira, aplicações que suportam apenas RDF podem entender pelo menos parcialmente um documento OIL.

Partindo-se do pressuposto que uma única linguagem de ontologias não pode se ajustar a todas as aplicações e usuários da Internet, OIL utiliza uma abordagem de camadas, cada uma adicionando funcionalidades e complexidade à camada inferior [33]. Assim, pessoas ou sistemas familiarizados apenas com as camadas mais inferiores podem entender parcialmente ontologias expressas em uma das camadas mais altas.

Uma ontologia em OIL é formada por duas seções: *ontology container* e *ontology definitions*. As Figuras 2.3 e 2.4 exibem exemplos, adaptados de [41], que ilustram cada uma destas seções.

A seção *ontology container* especifica algumas características gerais da ontologia, como nome, autor, assunto, linguagem, versão, etc. Para descrever estes atributos da ontologia utiliza-se o padrão de metadados Dublin Core [2]. Apesar de todo elemento neste padrão ser opcional, em OIL alguns são obrigatórios ou têm um valor pré-definido.

A segunda seção de uma ontologia OIL, *ontology definitions*, define a ontologia propriamente dita, a partir de um vocabulário ontológico particular. Ela inclui referências

ontology-definitions	value-type animal
slot-def come	class-def defined herbivoro
inverse e_comido_por	subclass-of animal
slot-def tem_como_parte	slot-constraint come
inverse e_parte_de	value-type planta OR
properties transitive	slot-constraint e_parte_de
class-def animal	has-value planta
class-def planta	class-def girafa
subclass-of NOT animal	subclass-of animal
class-def arvore	slot-constraint come
subclass-of planta	value-type folha
class-def ramo	class-def leao
slot-constraint e_parte_de	subclass-of animal
has-value arvore	slot-constraint come
class-def folha	value-type herbivoro
slot-constraint e_parte_de	class-def planta-gostosa
has-value ramo	subclass-of planta
class-def defined carnivoro	slot-constraint comido_por
subclass-of animal	has-value herbivoro, carnivoro
slot-constraint come	

Figura 2.4: Seção *ontology-definitions* de uma ontologia OIL.

a classes e listas de axiomas. A especificação atual da linguagem OIL ainda não define a estrutura dos axiomas. Para mais detalhes, consultar [38].

Uma outra iniciativa de linguagem para representação de conhecimento na Web é a DARPA Agent Markup Language (**DAML**) [26]. A motivação inicial para a criação desta linguagem foi a incapacidade de XML de expressar as relações existentes entre os conceitos representados nos seus documentos. Assim, DAML foi desenvolvida como uma extensão à XML e à RDF. A versão liberada em janeiro de 2001, **DAML+OIL**, provê meios para modelar domínios de conhecimento através de ontologias. DAML+OIL incorpora aspectos das linguagens DAML e OIL, e pode ser vista como um subdialeto desta última. Porém existem várias diferenças entre OIL e DAML+OIL, principalmente devido ao fato de que DAML+OIL foi baseada em RDF. Assim, algumas construções em RDF são possíveis em DAML+OIL mas não em OIL. Referências para esta questão podem ser encontradas em [26].

A linguagem **OWL** *Web Ontology Language* está sendo desenvolvida pelo Web Ontology Working Group do W3C, com o objetivo de prover meios para as aplicações entenderem o conteúdo das informações com as quais trabalham. Em OWL, termos de vocabulários podem ser representados explicitamente, bem como as relações entre entidades nestes vocabulários. Segundo o W3C, neste sentido OWL vai além de XML, RDF e

RDF Schema, ao permitir maior compreensão do conteúdo da Web. A linguagem OWL foi criada a partir de DAML+OIL, incorporando o que se aprendeu no desenvolvimento e aplicação de DAML+OIL. Documentação detalhada de OWL pode ser encontrada em [20].

2.4 Qualidade de dados

Um fator a ser considerado quando se trabalha com integração de dados é a importância de medidas de qualidade dos mesmos. O armazenamento de dados geográficos requer discretização da natureza, o que pode introduzir erros. Para que uma organização possa tomar decisões seguras, é necessário que os dados utilizados como suporte sejam precisos, atuais e consistentes. Além disto, deve-se armazenar aspectos relevantes sobre o dado para que o usuário possa formular julgamentos sobre a condição de seu uso. Orr [55] lista as seguintes regras gerais para qualidade de dados:

- dados não usados não permanecem corretos por muito tempo;
- a qualidade de dados em um Sistema de Informação é uma função de seu uso, não de sua coleção;
- a qualidade de dados não será, em última instância, melhor que seu uso mais rigoroso;
- problemas relativos à qualidade de dados tendem a piorar à medida que o sistema envelhece; e
- regras de qualidade de dados se aplicam tanto a dados quanto a metadados.

A quantificação da qualidade permite estabelecer critérios mínimos para integração de conjuntos de dados. Assim, indicadores de qualidade indicam quando é adequado intercambiar ou compartilhar dados e quais dados podem ser reusados [13]. Uma perturbação ocorre quando mais de uma fonte disponibiliza o mesmo dado, pois neste caso deve-se decidir qual dos valores disponíveis será utilizado e armazenado no sistema integrado. Nestas situações, indicadores de qualidade dos dados podem ser úteis para escolher a melhor fonte para cada dado, como proposto em [46].

Não há um único padrão estabelecido e utilizado em larga escala para avaliação de qualidade de dados. Na verdade, há vários padrões propostos, que se encaixam melhor em diferentes contextos. Não é simples estabelecer uma definição genérica de qualidade, já que só se deve avaliar a qualidade de um dado a partir da intenção de seu uso, em um

contexto específico. Parâmetros para a qualidade de dados geográficos podem ser obtidos em [48].

Avaliações objetivas de qualidade podem ser dependentes ou não da tarefa nas quais os dados são utilizados. Métricas *não-dependentes* da tarefa refletem características dos dados sem conhecimento contextual da aplicação, e podem ser aplicadas a qualquer conjunto de dados, não importando as tarefas a serem executadas. Métricas *dependentes* da tarefa devem ser desenvolvidas em contextos de aplicação específicos [56]. Podem incluir regras de negócio de organizações, regulamentos de governo ou de companhias, e restrições determinadas pelo administrador da base de dados.

Além disto, estudos têm confirmado que a qualidade dos dados é um conceito multi-dimensional. Neste contexto, *dimensões* são propriedades específicas de dados ou conjuntos de dados com propósitos de avaliação de qualidade, como, por exemplo, *completude*, *ausência de erro* e *atualidade*. Não existe na literatura um único conjunto de referência destas dimensões. Mecella et al. [46] propõem um conjunto de dimensões de qualidade composto por quatro propriedades: acurácia, completude, atualização e consistência interna. Já Pipino et al. [56] enumeram 16 dimensões diferentes, listadas na Tabela 2.1. O contexto da aplicação deve determinar, dentre estas dimensões, quais devem ser utilizadas.

Pode não haver uma definição clara e aceita de significado mesmo quando se trata de uma dimensão específica. Um exemplo é a dimensão de *completude*. Esta pode ser vista sob diferentes perspectivas, levando a diferentes métricas. No nível mais abstrato, por exemplo, completude está relacionada ao esquema, sendo um grau que indica a falta de entidades e atributos no esquema. No nível de dados, completude pode ser vista como a taxa de valores faltantes em uma coluna de uma tabela.

Uma dimensão de qualidade que merece atenção é a dimensão de *acurácia*. Em [58], acurácia está relacionada à proximidade entre o valor armazenado e o valor considerado correto. Esta definição é difundida na literatura e geralmente aceita. Desta maneira, para se medir a acurácia de um valor é necessário um oráculo, responsável por indicar qual é o valor correto. Nem sempre este oráculo se encontra disponível. Para alguns tipos de dados geográficos, como temperatura ou precipitação pluviométrica, podem ser aplicados métodos estatísticos, utilizando, por exemplo, medidas de datas próximas ou de localidades próximas, para estimar o valor considerado correto e compará-lo com o valor armazenado.

Têm sido propostas maneiras de combinar várias propriedades de qualidade em um único valor. Por exemplo, Mead [45] sugere uma métrica baseada na soma de 10 dimensões com pesos idênticos. Este cálculo pode não ser o mais eficaz, uma vez que atribui pesos idênticos a dimensões que medem características totalmente diferentes.

Costanza et al. [23] propõem um outro tipo de dimensão composta, calculada através de uma soma normalizada de três pontuações inteiras, de 0 a 4, representando avaliações

Dimensão	Definição
Acessibilidade	A medida de quanto os dados são acessíveis, ou fácil e rapidamente recuperáveis.
Atualidade	A medida de quanto os dados são suficientemente atuais para a tarefa em questão.
Ausência de erro	A medida de quanto os dados são corretos e confiáveis.
Completude	A medida de quanto os dados são completos.
Compreensibilidade	A medida da facilidade de compreender os dados.
Consistência de representação	A medida da uniformidade da representação dos dados.
Concisão de representação	A medida da compactação da representação dos dados.
Credibilidade	A medida de quanto os dados são tidos como verdadeiros e dignos de crédito.
Facilidade de manipulação	A medida de quanto os dados são fáceis de manipular e aplicáveis a tarefas diferentes.
Interpretabilidade	A medida de quanto os dados estão em linguagens, símbolos e unidades apropriados, e de quanto as definições estão claras.
Objetividade	A medida de quanto os dados são não-influenciados e imparciais.
Quantidade apropriada de dados	A medida de quanto o volume de dados é apropriado para a tarefa em questão.
Relevância	A medida da utilidade e aplicabilidade dos dados na tarefa em questão.
Reputação	A medida de quanto os dados são confiáveis em termos de sua origem.
Segurança	A medida de quanto o acesso aos dados é restrito apropriadamente para manter sua segurança.
Valor	A medida das vantagens e benefícios proporcionados pela utilização dos dados.

Tabela 2.1: Dimensões de qualidade de dados [56].

de três fatores: qualidade dos modelos, qualidade dos dados e grau de aceitação. Os pesos de cada pontuação são definidos de acordo com os critérios da aplicação. Mesmo assim, este índice composto pode misturar fatores com importâncias diferentes para usuários distintos.

Desta maneira, para realizar uma avaliação objetiva de qualidade, companhias e instituições devem seguir um conjunto de princípios para desenvolver métricas específicas para suas necessidades. Pipino et al. [56] propõem três formas funcionais de definição de métricas objetivas, descritas a seguir: proporção simples, operação mínima ou máxima e média ponderada.

A *proporção simples* expressa a razão entre o número de resultados desejados e o número total de resultados, normalizados entre 0 e 1 (valor ideal). Algumas métricas tradicionais de qualidade, como completude, consistência e ausência de erro podem ser avaliadas utilizando esta forma.

A *operação mínima ou máxima* pode ser utilizada quando se necessita agregar múltiplos indicadores de qualidade em uma única dimensão, normalizados entre 0 e 1. Computa-se o mínimo (ou máximo) valor dentre os valores normalizados de cada indicador de qualidade. No caso do operador mínimo, utiliza-se o menor valor normalizado para quantificar a qualidade da dimensão sendo avaliada. A *confiança*, que pode ser vista como a medida de quanto os dados são verdadeiros e confiáveis, é um exemplo de dimensão que pode fazer uso do operador mínimo. Ela pode refletir, por exemplo, uma combinação de três fatores: julgamento individual da credibilidade da fonte de dados, comparação com um padrão aceito e experiência prévia. Expressando cada uma destas três variáveis numa escala de 0 a 1, a confiança como um todo poderia ser expressada como o menor dos três valores.

O terceiro tipo de forma funcional de métricas de qualidade proposto em [56] é a *média ponderada*. É uma forma funcional que agrega múltiplos indicadores de qualidade em uma única dimensão. Neste caso, atribui-se um peso a cada uma das variáveis e calcula-se a média ponderada. Este tipo de cálculo só deve ser realizado se a organização tem um bom entendimento da importância de cada variável na avaliação geral da qualidade. Esta importância depende da aplicação.

Os processos pelos quais os dados geográficos passam apresentam diversas fontes de erros, que vão desde a etapa de coleta à etapa de apresentação ao usuário. Porém, Chrisman [16] defende que a responsabilidade sobre a qualidade dos resultados também deve ser atribuída ao usuário, uma vez que ele especifica várias das decisões de processamento e com isto determina muito do significado final.

No contexto de SIGs, pode fazer mais sentido aceitar que existem múltiplas perspectivas de “verdade”, ao invés de procurar especificar tudo com definições rígidas. Em análise ambiental deve-se esperar diferenças de opinião. Com isso fatores como acurácia

do dado não desaparecem, mas se tornam mais complicados. Poder-se-ia então ter várias perspectivas, cada uma sendo avaliada de acordo com o contexto de uso.

Assim como para dados convencionais, não há falta de padrões propostos para avaliação de dados espaciais. O problema hoje é escolher entre padrões incompatíveis. Padrões de acurácia para mapas foram definidos mesmo antes da era digital. Um exemplo é o US National Map Accuracy Standard (NMAS) de 1947, que define uma tolerância posicional onde 90% dos pontos devem estar dentro de 0,5mm de sua posição correta. Este padrão estabelece um critério para qualquer possível usuário. Este tipo de padrão, denominado *conformação com a expectativa*, pode ser adequado em períodos de tecnologia estável e produção centralizada, mas pode não ser adequado em situações nas quais as necessidades dos usuários não são conhecidas pelos produtores.

A questão de quando ou não adquirir ou utilizar uma base de dados está ligada a um julgamento que pode ser melhor definido como *adequação ao uso*: “esta fonte de informação serve suficientemente aos meus propósitos de maneira a compensar o esforço para obtê-la e convertê-la para o meu uso?”.

Desta visão geral, pode-se deduzir que qualidade de dados envolve vários aspectos. Especialmente, quando se trabalha com integração deve-se ponderar soluções de diferentes maneiras e perspectivas, o que requer um entendimento cuidadoso de cada fonte de dados e das aplicações nas quais os dados serão utilizados.

2.5 Séries históricas e dados climatológicos

Este trabalho está centrado em integração de dados climatológicos, que comumente são analisados na forma de séries históricas. Esta seção introduz o conceito de série histórica e descreve algumas das medidas climatológicas consideradas na arquitetura proposta.

Uma *série temporal* corresponde a um conjunto de valores onde cada valor é associado a uma marca de tempo. Uma mesma marca de tempo é associada a no máximo um valor. Desta maneira, o resultado é a variação dos valores em função do tempo. Séries temporais são utilizadas em diversos campos de estudo, como matemática e economia.

Quando uma série temporal contempla vários anos, ela pode ser considerada uma *série histórica*. No contexto desta dissertação, os dados manipulados consistem em séries históricas climatológicas. Tais séries são obtidas por *estações* climatológicas, que coletam e registram medidas utilizando uma determinada granularidade de tempo. Assim, tem-se o registro da variação de um aspecto climático com o passar do tempo. Como as séries são associadas à localização geográfica da estação onde foram coletadas, elas podem ser submetidas a processamentos geoestatísticos para extração de informações. Medidas climatológicas típicas incluem pressão atmosférica, umidade relativa do ar, radiação solar e precipitação.

A *pressão atmosférica* medida em um local corresponde à força por unidade de área exercida pelo peso do ar sobre a superfície do local em questão. A pressão atmosférica é medida com o uso de um instrumento chamado barômetro, e as medidas comumente utilizadas para registrá-la incluem milímetros de mercúrio (mmHg), kilopascal (kPa) e hectopascal (hPa). Apesar de ser determinada pelo peso do ar, a pressão em um mesmo local pode variar com o passar do tempo devido às mudanças de temperatura. Quando esta se eleva, o ar se expande e passa a ocupar uma área maior, resultando em uma diminuição da pressão. Quando a temperatura cai, ocorre o contrário: o ar se contrai, ocupando um espaço menor e provocando um aumento da pressão. Observações da variação da pressão atmosférica são bastante úteis em previsões de tempo.

A medida da *umidade do ar* indica a quantidade de vapor de água contida em um determinado volume de ar, e geralmente é expressada em gramas por metro cúbico (g/m^3). Porém, uma medida mais utilizada em análises climatológicas é o registro da *umidade relativa*, que indica a relação entre a medida da umidade do ar e a quantidade máxima de umidade que o ar poderia conter quando saturado. Assim, não se utiliza unidades para medir a umidade relativa, sendo esta expressada em porcentagem. Uma umidade relativa de 0% significa que o ar não possui nenhum vapor de água, e uma umidade relativa de 100%, que o ar não pode conter mais vapor de água. A umidade relativa pode aumentar, por exemplo, quando uma massa de ar movimenta-se sobre um corpo de água.

O sol, assim como todos os corpos, emite energia radiante. A medida da *radiação solar* corresponde à quantidade de energia recebida do sol na superfície terrestre, por unidade de área e de tempo. A radiação solar comumente é expressa em watts por metro quadrado (W/m^2) ou em calorias por centímetro quadrado por minuto ($\text{cal}/\text{cm}^2/\text{min}$). Embora a energia vinda do sol seja inesgotável e não ofereça riscos ambientais, ela ainda é pouco utilizada.

A próxima seção caracteriza dados pluviométricos, utilizados na implementação desta dissertação.

2.5.1 Dados pluviométricos

A quantidade de chuva que ocorre em um local, em uma determinada data, é expressada pela altura de água caída e acumulada sobre uma superfície plana e impermeável. Esta medida é efetuada em pontos previamente escolhidos, utilizando-se dois tipos de aparelhos: o *pluviômetro* e o *pluviógrafo*. O primeiro registra apenas a quantidade total de água precipitada, e o segundo registra também a intensidade da chuva. As medidas são realizadas periodicamente, geralmente em intervalos de vinte e quatro horas. No Brasil, as medidas normalmente são feitas às sete horas da manhã.

Apesar de haver vários tipos de pluviógrafos, somente três têm sido mais utilizados: o

Pluviógrafo de Caçambas Basculantes, o Pluviógrafo de Peso e o Pluviógrafo de Flutuador. O primeiro se baseia na oscilação de dois compartimentos arranjados de maneira que, a cada 0,25mm de água caída, um dos compartimentos bascula, acionando um registrador elétrico que registra a queda da água. O segundo se baseia em um sistema de pesos, no qual, à medida em que a água cai e é acumulada no receptor, uma pena é acionada traçando um gráfico sob a forma de um diagrama de massas, resultando na altura da precipitação acumulada *versus* tempo. O terceiro é bastante semelhante, com uma pena acoplada a um flutuador, de maneira que, à medida em que o nível de água do receptor sobe, a pena registra em um diagrama a altura da água caída com o decorrer do tempo. Acredita-se que este método de medição seja mais exato que o realizado pelos pluviógrafos de caçambas basculantes.

Desta maneira, os pluviógrafos são aparelhos mecânicos que produzem gráficos de precipitação. Erros de leitura, digitação ou transcrição de valores e ainda omissões do leitor do pluviômetro ou inoperância da estação por períodos de tempo podem fazer com que a medida da quantidade de chuva de uma data seja registrada errada, ou fazer com que uma série histórica apresente lacunas para determinados períodos de tempo.

Isto exige a aplicação de métodos de consistência sobre as séries históricas antes destas serem utilizadas. Estes métodos de consistência tentam detectar valores considerados duvidosos para a estação, e podem também tentar corrigir tais valores e preencher lacunas nas séries. A seguir são descritos alguns dos métodos de consistência apresentados na literatura.

Métodos de consistência temporal

Alguns métodos estimam dados faltantes através de médias aritméticas. Neste sentido, um método simples para obter o valor de uma medida faltante é calcular a média aritmética de datas anteriores e posteriores à data cujo valor está sendo estimado. Pode-se aplicar pesos às medidas, de maneira que datas mais distantes da data sendo estimada tenham pesos menores.

Um exemplo deste tipo de interpolação de série temporal é tomar como base as leituras anteriores e posteriores temporalmente mais próximas à estação em questão. Seja x a estação cuja medida se quer estimar, t_f a data do valor faltante, t_a e t_p respectivamente as mais próximas datas anterior e posterior à data do valor faltante. A precipitação $p_x(t_f)$ da estação x , para a data t_f , é dada pela interpolação:

$$p_x(t_f) = p_x(t_a) + \frac{t_f - t_a}{t_p - t_a} (p_x(t_p) - p_x(t_a)) \quad (2.1)$$

É importante notar que este tipo de método se baseia apenas no comportamento temporal, não se aproveitando do fato de que as medidas pluviométricas são georeferenciadas

e de que o aspecto da distribuição geográfica da chuva pode ser utilizado.

Um método bastante difundido, utilizado para análise de consistência de séries históricas pluviométricas é o método da *dupla-massa* [4]. Ele compara os valores acumulados anuais ou sazonais da estação x com uma estação de referência y . A estação de referência é usualmente uma estação virtual, cujas medidas podem ser determinadas a partir da média de medidas de diversas estações vizinhas. Os pares cumulativos (valores dupla-massa) são plotados em um plano cartesiano, e o gráfico é examinado à procura de mudanças de direção. Se o gráfico é essencialmente linear, os registros da estação x podem ser considerados consistentes, uma vez que eles correspondem aos registros da estação de referência. Se o gráfico mostra uma mudança de inclinação, os registros da estação x não são consistentes e pode-se neste caso multiplicar os valores posteriores à mudança de direção por um fator de correção. Este fator de correção é a razão entre o coeficiente angular da reta antes da mudança de direção e o coeficiente angular da reta após a mudança de direção. A Figura 2.5, retirada de [4], ilustra um exemplo no qual foi encontrada uma mudança de direção dos valores dupla-massa a partir da quinta medida acumulada.

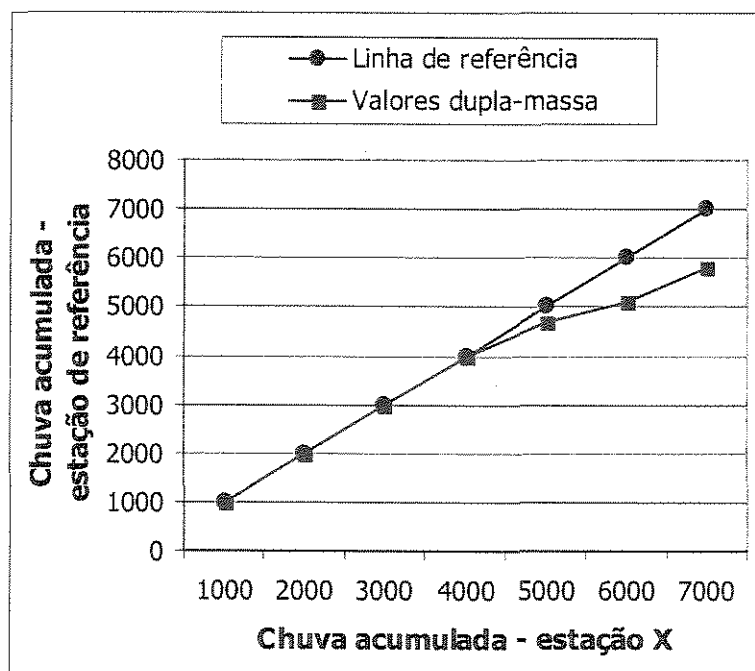


Figura 2.5: Exemplo de aplicação do método dupla-massa.

Métodos de consistência espacial

Enquanto alguns métodos de consistência levam em consideração exclusivamente o comportamento temporal da chuva, outros levam também em consideração o seu comportamento espacial. Estes utilizam como parâmetros medidas de estações geograficamente próximas à estação cujos dados estão sendo analisados ou estimados. De acordo com a topografia da área, métodos que consideram a característica espacial podem se mostrar bem mais eficazes do que aqueles que consideram apenas o aspecto temporal.

Um exemplo deste tipo de método, utilizado para obtenção de valores faltantes, é o *Método do Vizinho Mais Próximo* [44]. Neste método, uma lista de estações alternativas é associada à estação cujos valores estão sendo estimados. O valor faltante é tido como o valor da primeira estação da lista, na data em questão. Uma maneira simples de ordenar a lista de estações é basear-se na distância à estação sendo estimada.

Um exemplo de método de média aritmética que utiliza dados de estações vizinhas é o que toma como base as medidas pluviométricas de três estações localizadas o mais próximo possível da estação que apresenta a falta do dado, na data em questão. A medida faltante pode ser estimada pela média aritmética ponderada das medidas das três estações vizinhas, utilizando-se como peso as razões entre as precipitações médias anuais da estação em questão e de suas vizinhas. Sendo x a estação que apresenta a falta do dado e A, B, C as estações vizinhas, tem-se:

$$p_x(t_f) = \frac{1}{3} \left[\frac{N_x}{N_A} p_A(t_f) + \frac{N_x}{N_B} p_B(t_f) + \frac{N_x}{N_C} p_C(t_f) \right] \quad (2.2)$$

onde N é a precipitação média anual.

Outro método alternativo utiliza como base a precipitação de quatro estações A, B, C, D localizadas próximas à estação x de interesse, cada uma em um dos quatro quadrantes delimitados pelas linhas norte-sul e leste-oeste passando pela estação x . Estima-se a medida da estação x através da média ponderada das medidas das quatro estações vizinhas, onde o peso aplicado é inversamente proporcional ao quadrado da distância L de cada estação base à estação x . O cálculo é realizado utilizando a fórmula:

$$p_x(t_f) = \frac{\sum_{i=1}^4 \left[\frac{p_i(t_f)}{L_i^2} \right]}{\sum_{i=1}^4 \left(\frac{1}{L_i^2} \right)} \quad (2.3)$$

Estes métodos de estimativa de valores produzem apenas um valor para cada medida faltante. Entretanto, a possibilidade de se obter mais de um único valor para cada medida faltante pode ser considerada. Nesta caso a idéia geral é criar, para cada uma das n medidas faltantes, m possíveis valores alternativos, com m pequeno, e considerar que $n \cdot m$ diferentes conjuntos completos são possíveis.

Exemplos de outros métodos que também podem ser utilizados para análises de con-

sistência de séries históricas de medidas pluviométricas são a *Análise do Componente Principal* [43], *Time Interpolation of the Principal Component Scores Series* [44], *Penalty of the Principal Component Scores* [44] e *Método do Vetor Regional* [64].

2.6 Resumo

Este capítulo apresentou uma visão geral sobre os conceitos necessários ao entendimento da dissertação: o conceito e características de SIGs e de bancos de dados espaço-temporais; os problemas relacionados à integração de dados heterogêneos, e, particularmente, os problemas relacionados a dados geográficos; características das alternativas propostas hoje para intercâmbio de dados, incluindo metadados e ontologias; e características de dados climatológicos. Além disto o capítulo ainda apresentou questões relativas à avaliação de qualidade de dados.

O próximo capítulo discute os problemas envolvidos no projeto de um sistema de integração de dados climatológicos, e apresenta alguns requisitos de sistema coletados em um caso real.

Capítulo 3

Problemas e Requisitos do Sistema

O projeto de um sistema de integração de dados climatológicos envolve várias questões. Conflitos de heterogeneidade devem ser resolvidos para que os dados possam ser processados adequadamente. É necessário igualmente tratar problemas relativos à avaliação de qualidade dos dados: de que ponto de vista a qualidade deve ser avaliada e de que maneira ela deve ser medida. Além disto, também é preciso abordar questões relativas à maneira como os dados integrados devem ser disponibilizados para o usuário e, eventualmente, para outros sistemas. Este capítulo discute tais questões.

A solução proposta nesta dissertação foi validada através da implementação de um sistema de integração de dados pluviométricos desenvolvido na EMBRAPA. Este capítulo também apresenta os requisitos de usuário levantados na EMBRAPA, que serviram como uma das bases para a implementação.

A seção 3.1 discute de maneira geral os problemas envolvidos no projeto de um sistema de integração de dados climatológicos. Em seguida, a seção 3.2 apresenta os requisitos de usuário coletados na EMBRAPA. A seção 3.3 resume o capítulo.

3.1 Questões a serem consideradas

Esta seção resume de maneira geral as questões envolvidas no projeto de um sistema de integração de dados climatológicos, abordando questões relacionadas à heterogeneidade, avaliação de qualidade e disponibilização de dados.

3.1.1 Heterogeneidade

As instituições que coletam dados climatológicos no Brasil possuem métodos próprios de coleta e armazenamento de dados. Assim, não existe um padrão, difundido e aceito, de modelagem e disponibilização de dados. Cada instituição trabalha de maneira diferente,

e independente, das outras. Isto gera problemas de heterogeneidade em diversos aspectos.

O primeiro problema de heterogeneidade está relacionado ao fato de que as instituições não coletam os mesmos tipos de dados. Por exemplo, enquanto alguns órgãos coletam diariamente medidas pluviométricas e de temperatura, outros coletam apenas medidas pluviométricas.

Para um mesmo tipo de dado, a *unidade de medida* utilizada nem sempre é a mesma. A unidade utilizada para dados pluviométricos pode ser, por exemplo, milímetros ou décimos de milímetros. Um outro exemplo são coordenadas geográficas: algumas instituições fornecem latitude e longitude em graus, minutos e segundos; outras, em graus decimais. Outro problema são as diferenças de *precisão* encontradas para um mesmo tipo de dado. Tais diferenças podem ocorrer tanto em dados convencionais quanto em dados espaciais. No caso de dados convencionais é comum verificar diferenças no número de casas decimais entre dados fornecidos por fontes diferentes. Já no caso de dados espaciais, também é comum ocorrer diferenças de precisão em medidas de localização geográfica.

Um quarto tipo de problema quando se lida com integração de dados climatológicos ocorre devido ao fato de as fontes de dados utilizarem formatos de armazenamento diferentes. Os formatos tipicamente utilizados pelas fontes incluem, entre outros, arquivos texto, planilhas Excel e formato relacional em SGBDs.

Outro problema é o fato de uma parte dos dados fornecidos por algumas instituições ser *estimado*. Isto é, algumas instituições, além de fornecerem dados efetivamente coletados, fornecem também dados estimados a partir de dados coletados. Pode-se desejar tratar dados estimados e dados coletados de maneiras diferentes.

Além dos problemas relacionados aos dados, também ocorrem diferenças na periodicidade com a qual as fontes de dados disponibilizam seus dados. Enquanto algumas instituições liberam dados diariamente, outras fornecem grandes conjuntos de dados anualmente. Com isto, a maneira com que os dados de cada instituição devem ser gerenciados e integrados pode variar.

Resumindo, quando se trabalha com diversas fontes de dados climatológicos, podem ocorrer problemas de heterogeneidade em:

- tipos de dados e origem dos dados (dispositivos de coleta);
- unidades;
- precisão;
- formato de armazenamento;
- ocorrência e quantidade de dados estimados; e
- periodicidade de fornecimento de dados.

Qualquer abordagem de integração de dados climatológicos deve se preocupar em resolver tais problemas. Esta é a principal preocupação da arquitetura proposta.

3.1.2 Avaliação de qualidade

Como discutido no capítulo 2, a avaliação de qualidade de dados é um processo que envolve vários aspectos. Esta seção discute algumas das questões envolvidas na determinação de uma metodologia de avaliação de qualidade para um sistema de integração de dados climatológicos.

Escolha de indicadores

Embora não exista um padrão único de qualidade, existe consenso sobre alguns de seus parâmetros. Em qualquer avaliação de qualidade, é necessário estabelecer *indicadores*, isto é, as características que se deseja avaliar. Pode-se optar por avaliar qualidade utilizando vários indicadores e, para cada indicador, estabelecer métodos específicos de medição. Claramente, são preferíveis indicadores que possibilitem medidas quantitativas, mas isto nem sempre é possível.

A escolha dos indicadores a serem considerados deve levar em conta o contexto das aplicações que utilizarão os dados. De acordo com tal contexto, alguns indicadores podem se mostrar mais adequados que outros. Outro fator a ser considerado é a relevância para o usuário. Uma vez que um dos objetivos da avaliação de qualidade é fornecer informações a respeito dos dados que auxiliem tomadas de decisão, é importante que os indicadores escolhidos sejam relevantes para o usuário.

Definição de métricas

Escolhidos os indicadores de qualidade a serem utilizados, deve-se determinar a maneira com que cada indicador será medido. Idealmente deve-se associar aos indicadores métricas quantitativas. Vale notar que também na definição das métricas a serem utilizadas é importante considerar as necessidades do usuário. Preferencialmente as métricas escolhidas devem resultar em avaliações que forneçam ao usuário informações sobre a qualidade dos dados.

Para alguns indicadores, como *atualidade* e *completude*, não é difícil determinar uma métrica quantitativa. Porém a maioria dos indicadores é subjetiva e portanto mais difícil de medir. Indicadores qualitativos, em princípio, devem ser avaliados a partir de opiniões de especialistas que possuam intimidade com os dados. Pode-se, porém, tentar definir métricas que resultem em avaliações quantitativas para indicadores qualitativos. Um exemplo de indicador qualitativo de qualidade que poderia ser mensurado desta maneira

é o indicador *reputação*.

Determinadas as métricas para cada indicador de qualidade, surge a questão de estabelecer ou não uma métrica que combine em uma só medida as avaliações de todos os indicadores. Como mencionado no capítulo 2, geralmente isto é resolvido calculando-se uma média ponderada das medidas de todos os indicadores. Esta solução pode não ser a melhor, uma vez que para usuários diferentes indicadores distintos podem ter importâncias distintas. Além disto, um valor agregado pode fornecer menos informação que o conjunto de seus componentes isolados.

Definição de critérios de aceitação

Uma vez escolhidos os indicadores e determinadas suas métricas, pode-se estabelecer *padrões de aceitação* para cada indicador. Os padrões de aceitação correspondem aos valores mínimos considerados aceitáveis para cada indicador.

É importante notar que o próprio processo de conversão dos dados agrega erros adicionais aos dados. Assim, no momento da avaliação de qualidade, o dado pode apresentar dois erros combinados: o erro originado na fonte de dados e o erro introduzido no processo de conversão [24]. Isto deve ser levado em conta ao se estabelecer padrões de aceitação para as fontes de dados.

O padrão de aceitação pode ser definido para valores individuais, ou estabelecido como o número máximo de erros permitido para que um conjunto de dados seja aprovado. Neste caso é preferível que a verificação seja feita sobre todo o conjunto de dados, e não apenas sobre uma amostragem deste. Os conjuntos de dados que apresentarem um número de erros maior que o estabelecido pelo padrão de aceitação devem ser *reprovados*. Esta reprovação significa que houve falhas em etapas anteriores, seja na etapa de coleta ou conversão dos dados. O ideal é não haver nenhum erro, mas geralmente erros eventuais que não indicam uma tendência no conjunto de dados não são significativos. Todas estas considerações são válidas desde que seja possível determinar, dentro de um conjunto de dados, o que corresponde a um erro.

Existem três tipos de ação a tomar a partir da definição dos critérios de aceitação: somente armazenar dados que se encaixem nos critérios; armazenar todos os dados e deixar aos usuários finais a seleção daqueles que se encaixem dentro de critérios específicos; e uma combinação destas duas ações. Assim como cada indicador de qualidade pode ter importância distinta para usuários diferentes, cada usuário pode ter também o seu próprio critério de aceitação.

Uma vez que metadados associados aos dados também auxiliam no suporte a tomadas de decisão, também podem ser definidos critérios de aceitação sobre os metadados. Como o conjunto de metadados fornecidos pelas fontes de dados nem sempre é o mesmo, estes critérios podem determinar, por exemplo, conjuntos mínimos aceitáveis de metadados.

Neste caso, conjuntos de dados aos quais não estiverem associados tais conjuntos mínimos não seriam incorporados à base de dados.

Tratamento de séries temporais

Dados climatológicos geralmente se apresentam como séries temporais. Conjuntos de dados correspondentes a períodos de tempo são submetidos a processos estatísticos para extração de informações. Ou seja, frequentemente dados pontuais são menos relevantes do que quando vistos em conjunto.

Assim, em geral a qualidade de dados climatológicos é avaliada sobre séries temporais. Opcionalmente, uma abordagem que avalie tanto a qualidade de dados pontuais quanto de períodos de tempo pode ser utilizada.

Avaliar qualidade de séries temporais implica na escolha de indicadores, métricas e critérios de aceitação diferentes dos utilizados para dados pontuais. Uma questão específica consiste na determinação da granularidade de tempo a ser utilizada na avaliação de qualidade. Novamente as necessidades dos usuários podem ser consideradas para resolver estas questões.

Resultado da avaliação de qualidade

Este trabalho trata de integração de dados heterogêneos e de sua qualidade. No contexto de dados climatológicos, geralmente se trabalha com dados provenientes de diversas fontes (estações climatológicas). Neste caso, há duas opções a considerar:

- associar qualidade a cada fonte, integrar os dados provenientes das diferentes fontes e definir a qualidade do resultado final, mantendo porém informações sobre a qualidade de cada fonte original; ou
- integrar os dados e definir a qualidade do resultado da integração.

Como se verá, optou-se pela primeira opção. É importante manter a informação sobre a qualidade de cada fonte (as estações), para flexibilizar consultas sobre conjuntos de dados de fontes distintas.

Disponibilização dos resultados

Além de medir a qualidade dos dados, é necessário determinar uma maneira de explicitar claramente ao usuário os resultados do processo de avaliação de qualidade. Este problema pode ser mais facilmente solucionado se os indicadores e métricas de qualidade tiverem sido definidos de acordo com as necessidades dos usuários.

Além de informar ao usuário os resultados da avaliação de qualidade, é interessante estabelecer uma maneira de informar estes resultados também a qualquer aplicação que venha a fazer uso dos dados. Isto permite aos usuários destas aplicações avaliar a adequação dos dados ao uso pretendido.

Uma possível solução para este problema é disponibilizar os indicadores de qualidade dos dados sob a forma de metadados associados aos dados. Neste caso, é importante que tais metadados sejam disponibilizados em um formato legível por outros sistemas.

3.1.3 Disponibilização dos dados

Uma vez definida a forma de integração e de avaliação de qualidade dos dados, é necessário pensar na maneira como os estes serão disponibilizados. O primeiro destinatário dos dados no sistema considerado são seus usuários, que farão consultas sobre os dados. O segundo destinatário são outros sistemas.

Um maneira de facilitar tanto o acesso do usuário aos dados quanto a leitura dos mesmos por outros sistemas é utilizar o recurso de associar metadados aos dados, apresentado no capítulo 2. No contexto de dados climatológicos, o uso de metadados exige considerar algumas questões. Destacam-se as seguintes:

- definição do conjunto de atributos destes metadados;
- determinação da regra de codificação de cada metadado; e
- definição do formato de publicação dos metadados.

A definição do conjunto de atributos dos metadados a ser utilizado deve levar em consideração o escopo de dados climatológicos. Os metadados podem descrever, por exemplo, características relacionadas à *origem* e à *cobertura temporal* dos dados. Pode-se adotar um ou vários padrões de metadados propostos na literatura, como os apresentados no capítulo 2, para facilitar a aceitação dos metadados por outros usuários e sistemas. O mais importante é que os metadados descrevam características relevantes para o usuário, já que eles, juntamente com os resultados da avaliação de qualidade, devem fornecer informações suficientes para o usuário avaliar se o conjunto de dados é adequado aos seus propósitos. Além disto, a escolha dos metadados pode ser feita de maneira a permitir que eles sejam utilizados para auxiliar consultas sobre os dados. Por fim, deve-se tomar o cuidado de não prover uma quantidade demasiada de metadados ao usuário. É mais importante disponibilizar os metadados mais relevantes do que prover grande quantidade de metadados.

Definido o conjunto de metadados, deve-se determinar como cada metadado será codificado, já que um mesmo metadado pode ser representado de maneiras diferentes. Por

exemplo, um metadado que expressa a *cobertura espacial* de um conjunto de dados pode ser codificado através de um polígono, ou simplesmente pelas quatro coordenadas extremas (norte, sul, leste e oeste) da região a que se refere o dado. A regra de codificação de um metadado pode tomar a forma de vocabulários controlados, notações formais ou regras de *parsing*. Se a regra de codificação não for processável por outros sistemas, o valor do metadado ainda deve ser útil para um leitor humano. Existem na literatura algumas propostas de codificação e de vocabulários controlados que podem ser adotados, como o vocabulário de nomes geográficos do Getty Research Institute [39] e a recomendação do W3C para representação de datas e horas [67].

Para que os dados integrados e os metadados sejam facilmente aceitos por outros sistemas, é essencial que a sua *publicação* seja realizada utilizando um formato amplamente acessível e legível na Web. As duas principais opções são as linguagens XML e RDF, apresentadas no capítulo 2.

No contexto de dados climatológicos, deve-se decidir ainda se os metadados serão utilizados para descrever dados pontuais ou períodos de tempo. No caso de optar-se por períodos de tempo, pode-se fixar ou não uma granularidade de tempo a ser adotada.

Finalmente, deve-se decidir quais metadados serão obrigatórios e quais serão opcionais, e decidir se os metadados serão fornecidos anexados aos dados ou separadamente destes. Neste último caso, deve-se determinar como relacionar dados e metadados.

3.2 Requisitos do sistema

Esta seção apresenta os requisitos levantados na EMBRAPA para um sistema de integração de dados pluviométricos. Tais requisitos, juntamente com a solução proposta nesta dissertação, serviram como base para a implementação.

3.2.1 Visão geral

Atualmente diversos órgãos brasileiros coletam dados pluviométricos. Cada órgão possui seu próprio conjunto de estações pluviométricas e utiliza seus dados coletados isoladamente. A realização de consultas utilizando dados de vários órgãos em conjunto aumenta o potencial de obtenção de diversos tipos de informações. Além disto, quanto maior a quantidade de dados pluviométricos disponível sobre uma determinada região, melhor a caracterização climática da região.

O principal objetivo do sistema é permitir uma extração eficiente de informações climatológicas a partir de dados pluviométricos provenientes de diversas fontes heterogêneas. Assim, no momento do processamento de consultas, todos os dados devem ser vistos como um conjunto único. Alguns dos órgãos que pretendem fornecer dados ao sistema são o

Ministério da Agricultura, o Instituto Agrônomo de Campinas (IAC), o Departamento de Águas e Energia Elétrica do Estado de São Paulo (DAEE), o Instituto Nacional de Pesquisas Espaciais (INPE) e a Agência Nacional de Águas (ANA).

Alguns exemplos de informações que o sistema deve fornecer ao usuário:

- caracterização pluviométrica de estações e regiões;
- visualização em forma de gráficos de estatísticas de cada estação; e
- identificação de regiões com comportamento pluviométrico semelhante.

O sistema de integração de dados pluviométricos corresponde a um módulo de um sistema maior desenvolvido na EMBRAPA, o **Agritempo**, apresentado no capítulo 1. Dentro do Agritempo, o sistema de integração de dados pluviométricos tem o nome de *Módulo de Séries Históricas de Chuvas*.

Como o Agritempo será disponibilizado na Web, os usuários poderão ser quaisquer pessoas relacionadas à produção agrícola, como agricultores e pessoas ligadas a tomadas de decisão. Eventualmente o sistema poderá, inclusive, auxiliar responsáveis por decisões relacionadas à liberação de Crédito Rural e Seguro Rural.

O sistema possui dois tipos de acesso: restrito e irrestrito. Os dados são disponibilizados no acesso *irrestrito*. Qualquer pessoa pode utilizar dados e funcionalidades deste acesso. Para utilizar as funcionalidades do acesso *restrito* é necessário estar cadastrado no sistema. Neste acesso são disponibilizadas funcionalidades referentes à manutenção do banco de dados.

Para facilitar o entendimento, os requisitos de usuário foram divididos em duas partes, apresentadas nas seções a seguir: (1) Requisitos de pré-processamento, especificando os objetivos do pré-processamento de dados antes destes serem utilizados em consultas, e (2) Requisitos de consultas, definindo alguns dos produtos a serem disponibilizados ao usuário (dados, mapas e informações em geral).

Na ocasião da escrita deste texto ainda não haviam sido definidos todos os requisitos funcionais do Módulo de Séries Históricas de Chuvas. Posteriormente novas funcionalidades seriam acrescentadas. Apesar de tratar-se apenas de um módulo do sistema Agritempo, ao longo do texto utiliza-se o termo *sistema* para se referir ao Módulo de Séries Históricas de Chuvas.

3.2.2 Requisitos de pré-processamento

O sistema deve manter, além de medidas pluviométricas diárias, totais pluviométricos mensais de cada estação. Antes de serem armazenadas em formato digital, as medidas podem ser transmitidas à EMBRAPA por diversas formas, como fax ou telefone. Neste

processo podem ocorrer erros de leitura, digitação ou transcrição de valores, fazendo com que a medida da quantidade de chuva de uma data não seja armazenada corretamente na base de dados. Além disto omissões do leitor do pluviômetro e inoperância da estação por períodos de tempo podem provocar lacunas nas séries históricas. Neste contexto, *lacunas* correspondem a datas ou períodos de tempo cujas medidas não são informadas ao sistema.

Devido a tais fontes de erro, o sistema deve executar uma etapa de pré-processamento, que consiste em aplicar alguns métodos estatísticos e testes de integridade espaço-temporal aos dados. O pré-processamento deve ser realizado após os dados terem sido integrados e antes do processamento de consultas. Os objetivos desta etapa são:

- identificação de períodos de tempo, de cada estação, adequados para serem utilizados em consultas;
- detecção e marcação de dados diários duvidosos;
- estimativa de dados diários faltantes; e
- identificação e correção de totais mensais duvidosos.

Na identificação de períodos de tempo adequados para serem utilizados em consultas, as séries históricas devem ser analisadas quanto à sua extensão e quanto à existência de lacunas de dados. Nesta verificação, uma série pode ser inteiramente reprovada, ou ser aprovada mas ter alguns de seus anos reprovados. Os critérios definidos por climatologistas da Embrapa para aprovar ou não uma série e seus anos são os seguintes:

- a) Se a série como um todo possuir menos de 10 anos de extensão, e se o último ano da série for menor ou igual ao ano atual menos 3, considera-se que a série é demasiadamente curta e que a estação correspondente foi desativada. A série inteira é reprovada.
- b) Anos com pelo menos 60 dias consecutivos sem dados são reprovados.
- c) Anos que possuem pelo menos 60 dias consecutivos com quantidade de chuva registrada como zero no período de tempo considerado verão para a região onde se encontra a estação são reprovados. Para esta verificação o sistema deve manter um registro do período de regime de verão de cada estado do Brasil.

A detecção e marcação de dados diários duvidosos e a estimativa de dados diários faltantes deve ser implementada utilizando métodos já difundidos na literatura, como os apresentados na seção 2.5.1. Um *dado duvidoso* corresponde a uma medida que originalmente é informada ao sistema mas que evidencia problemas de confiabilidade após o processamento de métodos estatísticos. Tanto medidas diárias quanto totais pluviométricos

mensais podem ser considerados duvidosos. Após o pré-processamento da série histórica de uma estação, a classificação de seus dados diários pode ser de quatro tipos:

- a) dados originalmente informados ao sistema e que não foram considerados duvidosos;
- b) dados originalmente informados ao sistema e que foram considerados duvidosos, sendo marcados como tal;
- c) dados originalmente não informados ao sistema, que foram estimados e que não foram considerados duvidosos;
- d) dados originalmente não informados ao sistema, que foram estimados e que foram considerados duvidosos.

A identificação e correção de totais mensais duvidosos também deve utilizar métodos discutidos na literatura. Ao corrigir um total mensal, o sistema deve manter o total original na base de dados. O usuário deve ter a opção de utilizar totais mensais originais ou corrigidos no momento de processar consultas.

O controle de acesso ao Módulo de Séries Históricas de Chuvas é o mesmo do sistema Agritempo como um todo. Assim, há dois tipos de acesso: restrito e irrestrito. Todos os procedimentos de pré-processamento devem ser disponibilizados somente no acesso restrito, para que sejam executados apenas por pessoas autorizadas a atuar como curadores do banco de dados.

3.2.3 Requisitos de consultas

As consultas envolvendo dados pluviométricos podem considerar:

- a) valores e sua distribuição em períodos (como anos e meses);
- b) cálculos de médias e extremos;
- c) distribuição espacial para um determinado período; e
- d) variações espaço-temporais de valores ou extremos.

A especificação das consultas do tipo c) e d) não havia sido terminada no momento da escrita deste texto. Por este motivo apenas algumas das consultas do tipo a) e b) estão detalhadas conforme levantamento de requisitos do usuário.

O sistema deve exibir totais pluviométricos anuais por estação, médias dos totais e anos com extremos de totais. Esta exibição pode ser por valores ou em diagramas de barras. Neste último caso, o diagrama deve conter uma linha representando a média móvel de 10

anos. Juntamente com estas informações também deve ser exibida a distribuição da chuva ao longo do ano na estação correspondente. A Figura 3.1 ilustra um possível formato de saída deste tipo de consulta.

Além de médias e extremos para anos, também devem ser disponibilizadas consultas sobre outros períodos de tempo, a saber: meses, quinzenas, decêndios e pênadas. A Figura 3.2 exibe um exemplo onde são exibidas estatísticas calculadas a partir da escolha do mês de fevereiro de 1990 como período base.

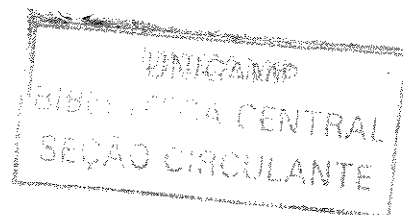
Caso o período de tempo escolhido pelo usuário contenha lacunas de dados, não deve ser exibido o total do período. Em vez disto, deve ser exibida ao usuário uma mensagem informando que aquele período não se encontra completo na base de dados. Porém as outras informações devem ser exibidas.

Todas as funcionalidades de consultas devem ser disponibilizadas no modo de acesso irrestrito. Além disto, no momento do processamento de consultas devem ser informadas ao usuário informações sobre a qualidade dos dados.

3.3 **Resumo**

Este capítulo descreveu as principais questões envolvidas no projeto de um sistema de integração de dados pluviométricos: questões relativas à resolução de conflitos de heterogeneidade, questões relativas ao estabelecimento de um método de avaliação de qualidade de dados, e questões relativas à disponibilização dos dados ao usuário e a outros sistemas. O capítulo também apresentou alguns dos requisitos de sistema coletados na EMBRAPA para o desenvolvimento de um sistema de integração de dados pluviométricos.

O próximo capítulo apresenta a solução proposta para o projeto de um sistema de integração de dados climatológicos heterogêneos.



Estação: Carazinho

Período solicitado: ano de 1998

	ANO	TOTAL (mm)
Total anual solicitado	1998	24332
Média anual	1941 a 1998	15687
Maior total anual	1990	27687
Menor total anual	1952	8101

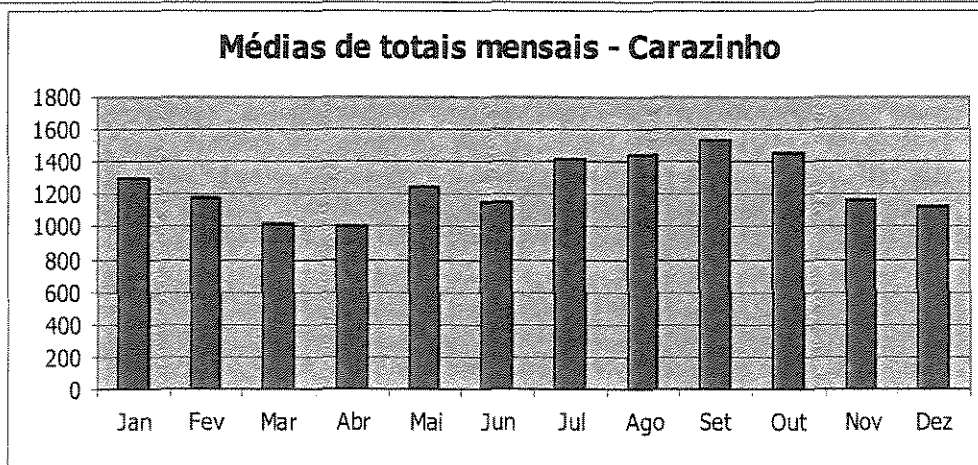
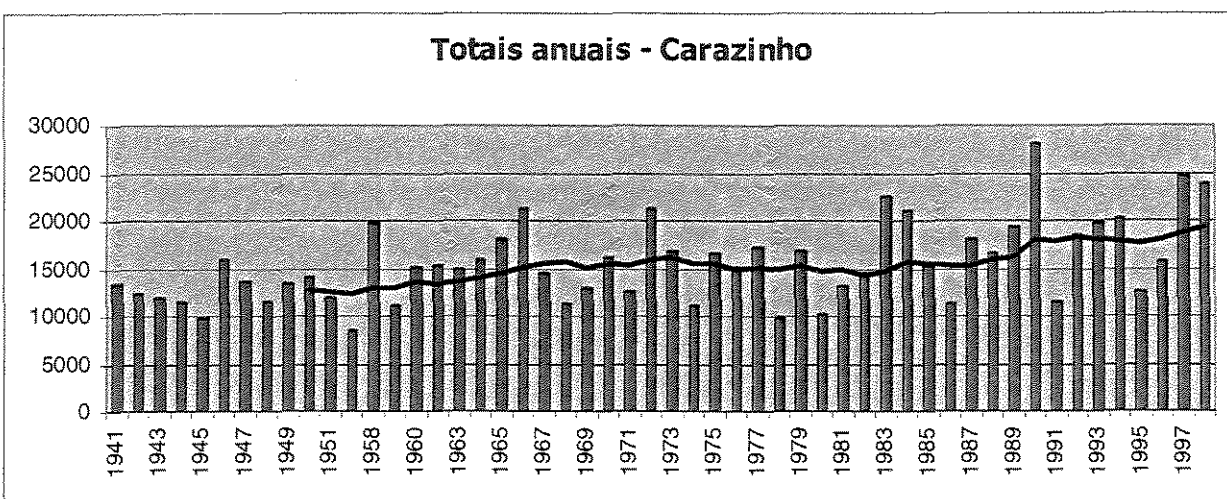


Figura 3.1: Exemplo de estatísticas a serem exibidas para anos.

Estação: Capanema

Período solicitado: mês de fevereiro de 1990

	ANO	TOTAL (mm)
Total do mês solicitado	1990	253
Média do mês	1977 a 1998	349
Maior total do mês	1985	462
Menor total do mês	1983	243

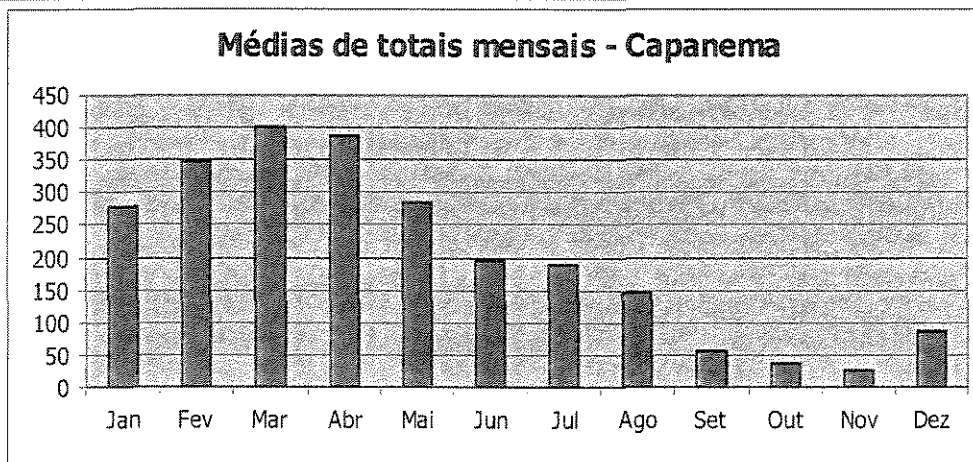
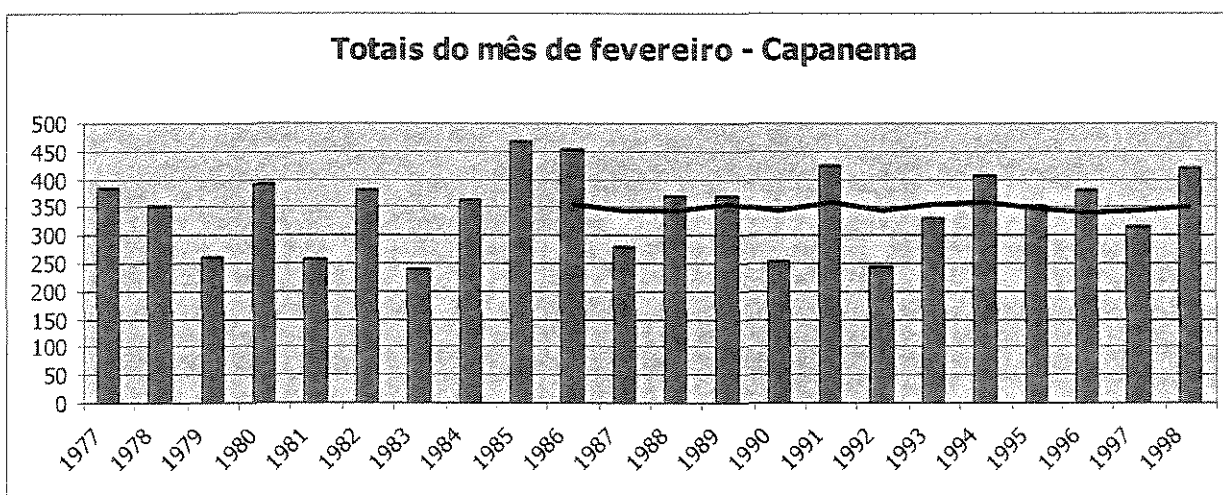


Figura 3.2: Exemplo de estatísticas a serem exibidas para meses.

Capítulo 4

Arquitetura Proposta

Este capítulo apresenta a arquitetura proposta para um sistema de integração de dados climatológicos. A solução procura resolver os problemas e requisitos de usuário discutidos no capítulo 3. Para tanto, a arquitetura aborda, entre outras, as questões de homogeneização de dados e de avaliação de qualidade.

A seção 4.1 fornece uma visão geral da arquitetura do sistema. A seção 4.2 descreve o banco de dados do sistema integrado. A seção 4.3 apresenta o banco de metadados utilizado para descrever os dados. As três seções seguintes descrevem os módulos do sistema, a saber: a seção 4.4 descreve o Módulo de Integração de Dados, a seção 4.5, o Módulo de Avaliação de Qualidade, e a seção 4.6, o Módulo de Processamento de Consultas.

4.1 Arquitetura do sistema

A solução proposta é baseada em um sistema centralizado. Os dados heterogêneos recebidos das fontes são inicialmente padronizados, integrados em um banco de dados com metadados associados e a seguir submetidos a um processo de avaliação de qualidade. Dados, metadados e indicadores de qualidade são armazenados em formato relacional em um SGBD.

A Figura 4.1 apresenta um diagrama simplificado do sistema, esquematizado em três módulos: **Módulo de Integração de Dados**, **Módulo de Avaliação de Qualidade** e **Módulo de Processamento de Consultas**.

Módulo de Integração de Dados - Este módulo é responsável por receber os dados de cada fonte, convertê-los para um formato homogêneo e inseri-los na base de dados integrada. A função do Módulo de Integração de Dados é assim resolver problemas de heterogeneidade, como diferenças de formato, esquema, unidades, precisão, e outros como

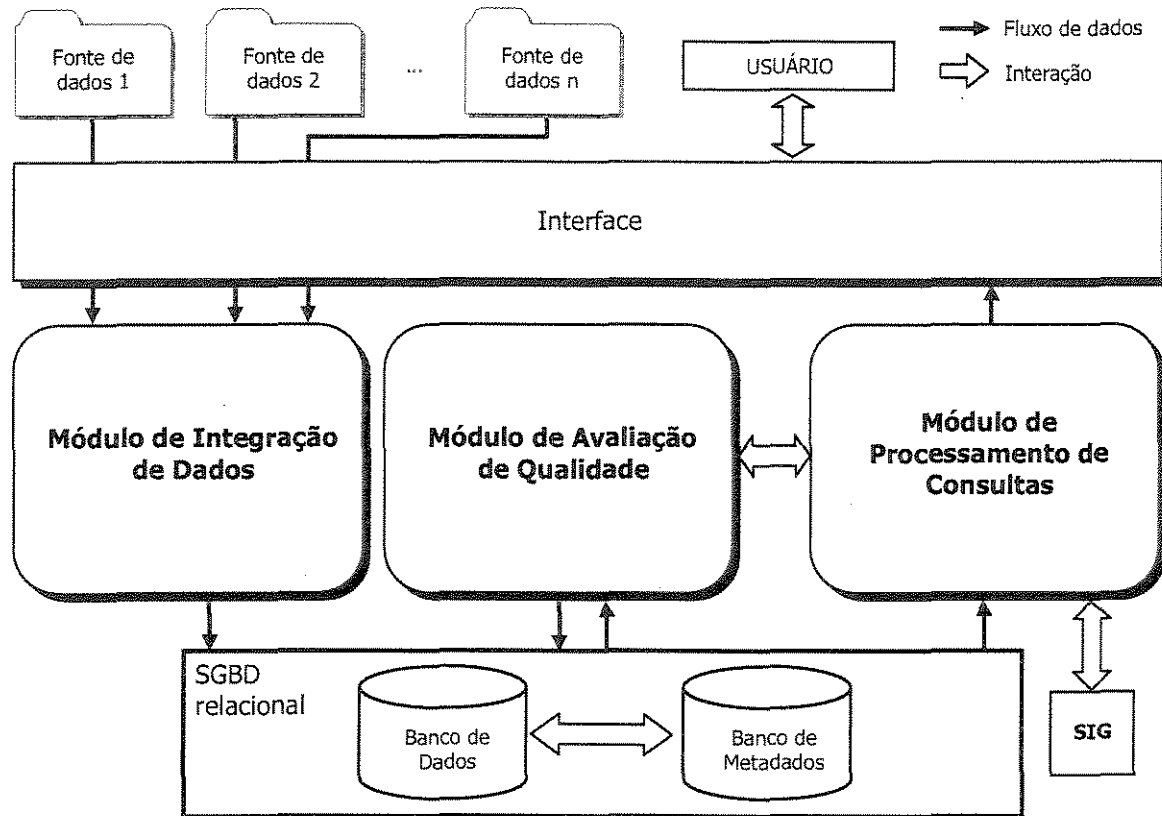


Figura 4.1: Diagrama da arquitetura proposta.

os discutidos no capítulo 3. O formato final de integração é aquele definido pelas necessidades das aplicações-alvo da EMBRAPA.

Módulo de Avaliação de Qualidade - Este módulo é responsável por avaliar a qualidade dos dados armazenados na base integrada, baseando-se na discussão do capítulo 3. Os resultados das avaliações expressam a qualidade dos dados de maneira a permitir ao usuário determinar que conjuntos de dados servem ao uso pretendido, e fornecem parâmetros da credibilidade para o resultado das consultas.

Módulo de Processamento de Consultas - Este módulo é responsável por processar consultas sobre o banco de dados integrado, utilizando também o banco de metadados e os resultados da avaliação de qualidade.

4.2 Banco de Dados

O banco de dados armazena de forma integrada os dados climatológicos heterogêneos provenientes das fontes de dados, segundo um *esquema global* que integra os esquemas de cada fonte de dados.

Medidas climatológicas consideradas

Com exceção da temperatura, todos os tipos de medidas climatológicas consideradas foram introduzidos no capítulo 2. As medidas climatológicas envolvem os seguintes fatores:

- temperatura;
- umidade relativa;
- precipitação;
- radiação solar; e
- pressão atmosférica.

Unidades e precisões de medidas

A Tabela 4.1 exibe as **unidades** e **precisões** adotadas no esquema global para cada tipo de medida climatológica considerada. Como mencionado no capítulo 2, a umidade relativa não possui unidade de medida própria, sendo expressada em porcentagem.

Tipo de medida	Unidade	Precisão
Latitude e longitude	Graus decimais	2 casas
Precipitação	Milímetros	1 casa
Temperatura	Graus Celsius	1 casa
Pressão atmosférica	Hectopascal	0 casas
Umidade relativa	Porcentagem	1 casa
Radiação solar	Calorias por centímetro quadrado por minuto	0 casas

Tabela 4.1: Unidades de medidas e precisões do esquema global.

Granularidades de tempo

O esquema global suporta dois tipos de granularidade de tempo: horária e diária. A **temperatura** pode ser registrada com as duas granularidades. Medidas de temperatura associadas a uma hora de uma data correspondem à temperatura no momento em questão, sendo armazenadas com granularidade horária. As **temperaturas mínima** e **máxima** de cada data são registradas com granularidade diária. A **umidade relativa** também pode ser registrada com as duas granularidades, sendo armazenadas com granularidade diária as **umidades relativas máxima, mínima e média** de cada data.

A **pressão atmosférica** pode ser registrada apenas com granularidade horária. A **precipitação acumulada** pode ser registrada com as duas granularidades: a precipitação acumulada associada a uma hora de uma data corresponde ao total de chuva na data até o momento em questão (granularidade horária), e a precipitação acumulada associada a uma data corresponde ao total de chuva ocorrida na data em questão (granularidade diária).

Por fim, a **radiação solar** também pode ser registrada com as duas granularidades. Com granularidade horária registra-se a radiação corrente em um determinado momento, e com granularidade diária armazena-se a **radiação solar acumulada**, que representa o total de energia recebida do sol, por unidade de área, em uma determinada data. A Tabela 4.2 resume as granularidades de tempo do esquema global.

Tipo de medida	Granularidade(s)
Temperatura	horária
Temperatura mínima / máxima	diária
Umidade relativa	horária
Umidade relativa mínima / média / máxima	diária
Precipitação acumulada	diária e horária
Radiação solar	horária
Radiação solar acumulada	diária
Pressão atmosférica	horária

Tabela 4.2: Granularidades de tempo para cada tipo de medida no esquema global.

Registro de estações vizinhas

O banco de dados armazena explicitamente as *estações vizinhas* de cada estação climatológica. Uma estação é considerada vizinha de outra se a distância entre elas é menor que um determinado valor e se elas se encontram numa mesma faixa de altitude. No momento

da inserção de uma estação, sua localização e altitude são comparadas com a localização e altitude de todas as estações já existentes no banco de dados, sendo armazenado um novo registro para cada estação considerada vizinha.

O propósito de armazenar explicitamente a vizinhança é permitir agilizar a aplicação de métodos de consistência nas séries históricas climatológicas. Como apresentado no capítulo 2, estes métodos permitem estimar medidas faltantes ou questionar medidas quanto à sua correteza. Em ambos os casos, geralmente utiliza-se medidas de estações vizinhas à estação considerada. Os critérios de distância e de faixas de altitude são definidos de acordo com os métodos a serem aplicados sobre as estações vizinhas. A distância entre estações vizinhas é armazenada explicitamente, para facilitar os métodos de consistência.

A Figura 4.2 exibe o diagrama entidade-relacionamento do esquema global. Como mostra a figura, no esquema global, além de medidas climatológicas são armazenados dados sobre as fontes de dados (as instituições), as estações climatológicas, suas vizinhas e seus municípios. Cada série histórica possui *anos* para cada tipo de medida climatológica. Os anos recebem notas para dois indicadores de qualidade, como detalhado mais adiante na seção 4.5. Não é obrigatório que os registros de granularidade horária e diária contenham medidas de todos os cinco tipos de medidas climatológicas consideradas.

4.3 Banco de Metadados

Os dados do banco de dados integrado são descritos por um *banco de metadados*, gerenciado pelo mesmo SGBD. O propósito do banco de metadados é ajudar o usuário a acessar os conjuntos de dados de interesse, e auxiliar consultas sobre os dados. Além disto os metadados facilitam a interoperabilidade caso os dados do sistema integrado venham a ser fornecidos para outros sistemas. Para tanto, os metadados devem descrever características que comumente são causas de problemas de heterogeneidade entre conjuntos de dados climatológicos. Visando atingir estes objetivos, propõe-se que o banco de metadados descreva as seguintes características dos dados:

- identificação;
- cobertura espacial;
- cobertura temporal;
- unidades utilizadas;
- administração dos metadados.

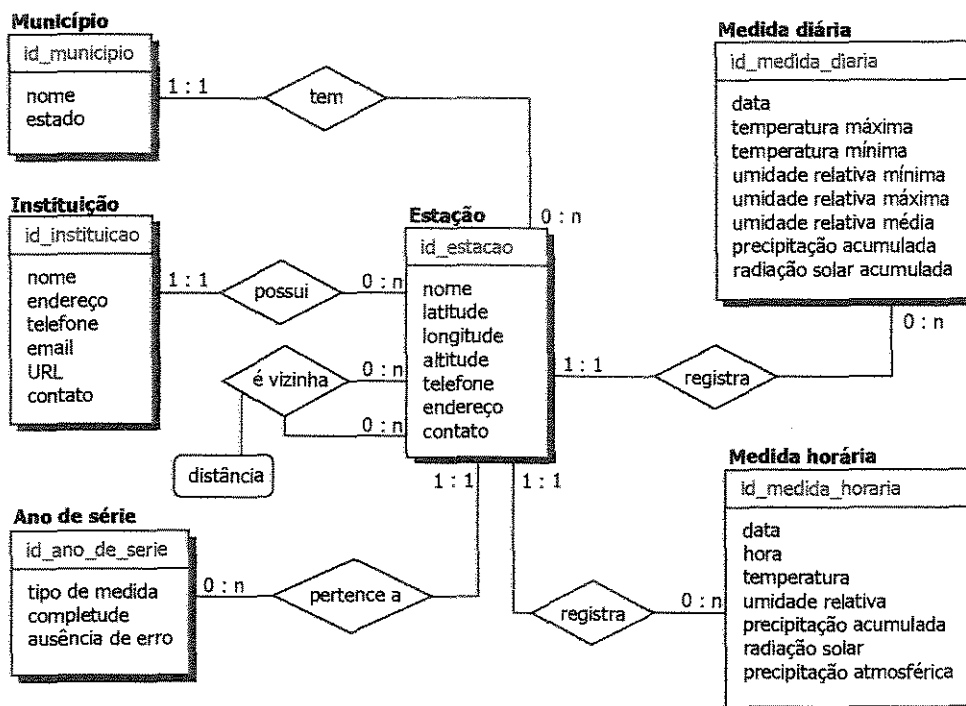


Figura 4.2: Diagrama entidade-relacionamento do esquema de dados global.

Todos os metadados, com exceção daqueles relacionados à qualidade, devem ser fornecidos pelas fontes juntamente com os dados. Adotou-se a convenção de utilizar o termo *elemento* para designar cada atributo do banco de metadados. Visando facilitar a interoperabilidade, os nomes de cada elemento são definidos em inglês. O Banco de Metadados é descrito nas seções 4.3.3 a 4.3.7.

É importante notar que o conjunto de metadados proposto pode ser utilizado por qualquer sistema que manipule dados climatológicos, mesmo que não se adote toda a arquitetura proposta.

4.3.1 Relacionamento entre dados e metadados

O banco de metadados armazena várias *descrições*. Cada descrição caracteriza um *conjunto de dados* de uma determinada fonte. Desta maneira, os dados fornecidos por uma fonte podem ter várias descrições no banco de metadados, uma para cada conjunto de dados com características próprias. Um conjunto de dados pode ser constituído, por exemplo, pelas medidas de um ano de uma série histórica, pelas medidas de toda uma

série histórica, ou por todas as medidas de um conjunto de estações.

Como mencionado, as descrições do banco de metadados devem ser fornecidas pelas fontes de dados quando os dados chegam ao sistema. Cada fonte determina como organizar seus dados em conjuntos a serem descritos pelos metadados.

A Figura 4.3 exibe como dados e metadados se relacionam. No banco de metadados, a cada descrição é atribuído um identificador. No banco de dados, cada registro de medida horária e diária é associado a um identificador de descrição do banco de metadados.

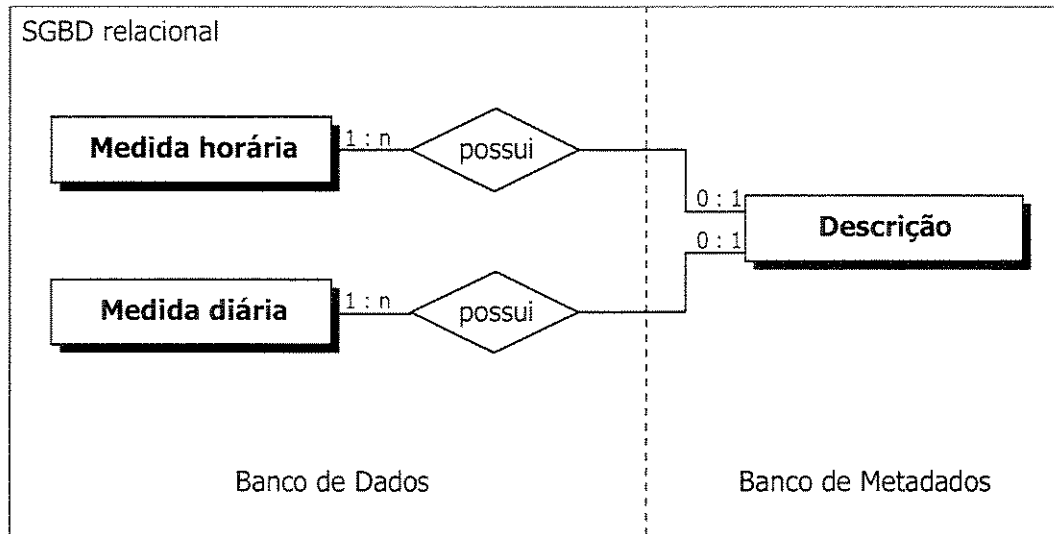


Figura 4.3: Relacionamento entre dados e metadados.

Por exemplo, considere que a instituição ANA é responsável pelas medidas climatológicas da estação de Ribeirão Preto. Pode-se estabelecer que a descrição do conjunto de dados da estação de Ribeirão Preto seja associada ao identificador desc.ANA.ribeiraoPreto. No banco de dados, o conjunto de dados climatológicos seria decomposto em vários registros de acordo com a natureza e a granularidade temporal. Cada registro de medida diária e horária desta estação teria este valor como um de seus atributos, permitindo consultar sua descrição no banco de metadados.

4.3.2 Padrões utilizados

A elaboração do conjunto de metadados foi realizada a partir das seguintes propostas de metadados existentes na literatura (algumas já introduzidas no capítulo 2):

- *Content Standards for Digital Geospatial Metadata (CSDGM)* [30], do Federal Geographic Data Committee (FGDC);
- *WMO Core Metadata Standard* [54], da World Meteorological Organization (WMO);

- *Metadados para Workflows Científicos no Apoio ao Planejamento Ambiental* [59]; e
- *Dublin Core Metadata Initiative (DC)* [2].

Os três primeiros foram considerados por descreverem dados cujo escopo se relaciona com o escopo de dados climatológicos. Já o padrão Dublin Core, apesar de não ser voltado a nenhuma área específica de aplicação, foi utilizado para direcionar a escolha do nome de cada elemento do banco de metadados. Isto facilita a interoperabilidade e a interpretação do banco de metadados, uma vez que a semântica de cada termo do padrão DC está se tornando um padrão mundial. Assim, foram aproveitados deste padrão termos como `Rights`, `Date.issued` e `Coverage`.

Além dos padrões de metadados citados, também foram adotadas algumas convenções de *codificação* de metadados propostas na literatura, apresentadas ao longo das próximas seções. A adoção de padrões de codificação e de vocabulários controlados diminui a heterogeneidade e facilita a aceitação dos metadados por outros sistemas.

Os elementos do banco de metadados são divididos em cinco partes: Metadados de Identificação, Metadados de Cobertura Espacial, Metadados de Cobertura Temporal, Metadados de Unidades e Metadados Administrativos.

4.3.3 Metadados de Identificação

Os **Metadados de Identificação** descrevem características básicas do conjunto de dados, incluindo informações a respeito de seu conteúdo, propósito e estado, e possibilitam sua indexação. A Tabela 4.3 exibe os Metadados de Identificação. Nesta tabela, a coluna *Ocorrências* especifica o número permitido de ocorrências de cada elemento, indicando, conseqüentemente, quando um elemento é opcional ou obrigatório. A coluna *Formato / domínio* especifica o formato do elemento, ou, se for o caso, um conjunto de valores que preferencialmente deve ser seguido.

A codificação do elemento `Date.modified`, assim como a de todos os outros elementos que envolvem datas, segue a especificação do W3C para formatos de data e hora [67]. Esta especificação foi baseada no padrão ISO 8601, no qual uma data com os valores do dia, mês e ano é expressada no formato YYYY-MM-DD.

4.3.4 Metadados de Cobertura Espacial

Os **Metadados de Cobertura Espacial** descrevem propriedades espaciais do conjunto de dados. Eles permitem a realização de consultas sobre a região geográfica a que os dados se referem. A Tabela 4.4 exibe os Metadados de Cobertura Espacial.

Os elementos `Latitude` e `Longitude` são codificados segundo as convenções do padrão CSDGM do FGDC. Latitudes e longitudes são expressadas em graus decimais, com a

Elemento	Descrição	Ocor- rências	Formato / domínio
Identifier	Identificador do conjunto de dados.	1..1	<i>texto</i>
Title	Título do conjunto de dados.	1..1	<i>texto</i>
Description	Descrição do conjunto de dados.	1..1	<i>texto</i>
Purpose	Propósito do conjunto de dados e suas limitações.	1..1	<i>texto</i>
Status	O estado do conjunto de dados.	1..1	<i>“complete”, “incomplete”, “in modification”</i>
Update frequency	A frequência com que adições e modificações são feitas sobre o conjunto de dados.	1..1	<i>“daily”, “monthly”, “yearly”, “irregularly”</i>
Rights	Restrições e pré-requisitos legais para utilizar os dados.	0..1	<i>texto</i>
Creator	Pessoa ou organização que produziu os dados.	1..n	<i>texto</i>
Address	Forma de contato com o produtor dos dados.	0..n	<i>texto</i>
Date.modified	Data da última modificação realizada sobre o conjunto de dados.	0..1	<i>data</i>
Lineage	Descreve a fonte de dados, as fontes auxiliares, as transformações ocorridas e outros processamentos realizados sobre o conjunto de dados.	0..1	<i>texto</i>
Format	Formato em que se encontra o conjunto de dados.	1..1	<i>“relational”, “ASCII”, “Excel”, “other”</i>
Keywords	Palavras que resumizam o conteúdo do conjunto de dados; podem ser utilizadas para indexação.	1..n	<i>texto</i>
Relation.isPartOf	Conjunto de dados maior do qual o conjunto de dados faz parte.	0..1	<i>texto</i>
Online linkage	URL de página Web onde o conjunto de dados pode ser obtido.	0..1	<i>URL</i>

Tabela 4.3: Metadados de Identificação.

Elemento	Descrição	Ocor- rências	Formato / domínio
Region	Região referenciada pelo conjunto de dados.	0..n	<i>texto</i>
Location	Município referenciado pelo conjunto de dados.	1..n	<i>(município, estado, país)</i>
Latitude	Latitude do município.	1..n	<i>número (grau decimal)</i>
Longitude	Longitude do município.	1..n	<i>número (grau decimal)</i>

Tabela 4.4: Metadados de Cobertura Espacial.

precisão desejada. Latitudes ao norte do e sobre o Equador são precedidas por um sinal positivo (+) ou por nenhum sinal, e latitudes ao sul do Equador são precedidas por um sinal negativo (-). Longitudes a leste do ou sobre o meridiano primário são precedidas por um sinal positivo (+) ou por nenhum sinal, e longitudes a oeste do meridiano primário são precedidas por um sinal negativo (-). Estas convenções são compatíveis com a especificação ANSI X3.61 de representação de pontos geográficos para intercâmbio de informações.

4.3.5 Metadados de Cobertura Temporal

Os **Metadados de Cobertura Temporal** descrevem características temporais do conjunto de dados. Eles possibilitam consultas sobre a granularidade e períodos de tempo aos quais os dados se referem. A Tabela 4.5 exibe os Metadados de Cobertura Temporal.

Elemento	Descrição	Ocor- rências	Formato / domínio
Coverage.temporal	Cobertura temporal do conjunto de dados.	1..n	<i>(data de início, data de fim)</i>
Granularity	Granularidade de tempo com a qual há dados climatológicos no conjunto de dados.	1..1	<i>“monthly”, “daily”, “hourly”, “irregular”</i>

Tabela 4.5: Metadados de Cobertura Temporal.

4.3.6 Metadados de Unidades

Os **Metadados de Unidades** informam as unidades de medidas utilizadas no conjunto de dados. Eles incluem unidades de medidas de dados espaciais e convencionais, e a precisão de medidas de localização geográfica. A Tabela 4.6 exibe os Metadados de Unidades.

Elemento	Descrição	Ocor- rências	Formato / domínio
Latitude unit	Unidade utilizada em medidas de latitude.	1..1	<i>“decimal degrees”, “decimal seconds”, “degress, minutes and seconds”</i>
Longitude unit	Unidade utilizada em medidas de longitude.	1..1	<i>“decimal degrees”, “decimal seconds”, “degress, minutes and seconds”</i>
Latitude resolution	A diferença mínima entre dois valores de latitude adjacentes.	1..1	<i>número</i>
Longitude resolution	A diferença mínima entre dois valores de longitude adjacentes.	1..1	<i>número</i>
Altitude unit	Unidade utilizada em medidas de altitude.	0..1	<i>“meters”, “decimeters”</i>
Precipitation unit	Unidade utilizada em medidas de precipitação.	0..1	<i>“millimeters”, “tenths of millimeters”</i>
Temperature unit	Unidade utilizada em medidas de temperatura.	0..1	<i>“degrees Celsius”, “degress Fahrenheit”</i>
Atmospheric pressure unit	Unidade utilizada em medidas de pressão atmosférica.	0..1	<i>“millimeters of mercury”, “hectopascal”, “kilopascal”</i>
Relative humidity unit	Unidade utilizada em medidas da umidade relativa do ar.	0..1	<i>texto</i>
Radiation unit	Unidade utilizada em medidas da radiação solar.	0..1	<i>“watts by square meter”, “calories by square centimeter by minute”</i>

Tabela 4.6: Metadados de Unidades.

4.3.7 Metadados Administrativos

Os **Metadados Administrativos** apresentam informações sobre os próprios metadados, como a data da última revisão e a pessoa responsável pela manutenção dos metadados. A Tabela 4.7 exibe os Metadados Administrativos.

Elemento	Descrição	Ocor- rências	Formato / domínio
Responsible	Pessoa responsável pela administração dos metadados.	1..1	<i>texto</i>
Contact	Meios de contato com o responsável.	0..1	<i>texto</i>
Language	Linguagem utilizada nos metadados.	1..1	<i>texto</i>
Last revision	Data da última revisão dos metadados.	0..1	<i>data</i>
Next revision	Previsão da próxima revisão dos metadados.	0..1	<i>data</i>

Tabela 4.7: Metadados Administrativos.

O elemento *Language* é codificado segundo o padrão RFC 3066 [6], que, em conjunto com o padrão ISO 639, define etiquetas de duas e três letras, com possíveis sub-etiquetas, para a codificação de línguas. Exemplos incluem “pt” para português e “en” ou “eng” para inglês. Este padrão é largamente utilizado na Web e recomendado pelo Dublin Core.

O Apêndice A apresenta o esquema resultante da especificação do conjunto de metadados.

4.4 Módulo de Integração de Dados

De maneira geral, a função deste módulo é integrar os dados heterogêneos das fontes de dados em uma base homogênea. Para tanto, o Módulo de Integração de Dados realiza as seguintes ações:

1. Recebimento e leitura dos dados de cada fonte, nos seus formatos e esquemas originais.
2. Conversão dos dados para um formato e esquema globais.
3. Resolução dos problemas relativos à integração.
4. Inserção dos dados convertidos na base de dados integrada.

A leitura e conversão dos dados de cada fonte são feitas por *tradutores de dados*. Após convertidos, os dados são submetidos a um *migrador*, que os insere na base de dados. A Figura 4.4 exhibe o Módulo de Integração de Dados de maneira esquemática. As duas próximas seções detalham os tradutores e o migrador de dados.

Vale ressaltar que no domínio estudado, tipicamente, cada estação é específica de uma instituição. No entanto, há possibilidade de envio de dados sobre uma estação por

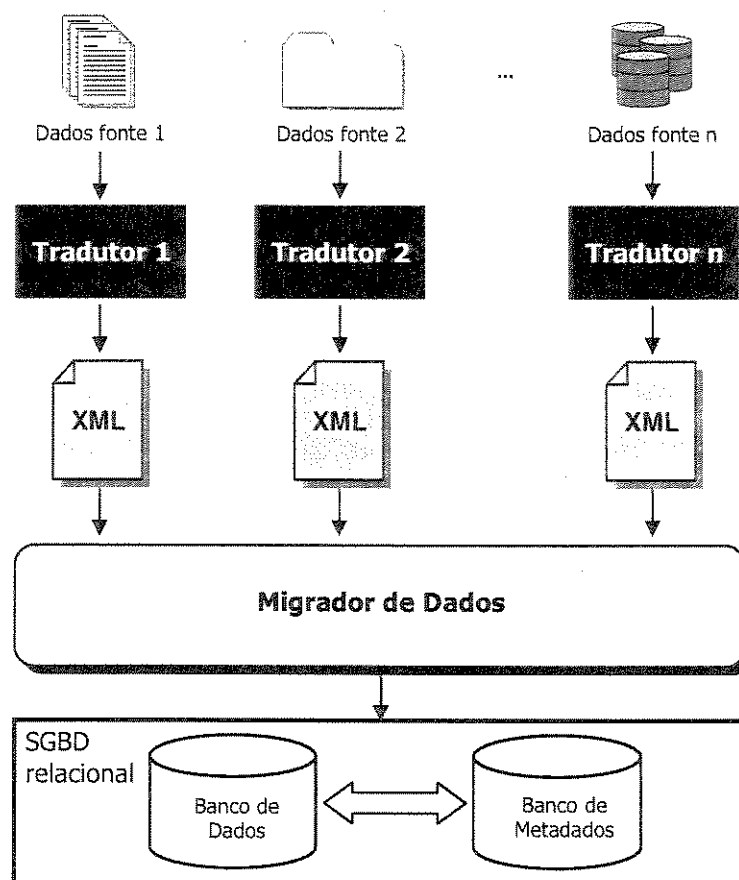


Figura 4.4: Módulo de Integração de Dados.

instituições diferentes. Isto ocorre, por exemplo, quando uma instituição reusa dados de outra. Outro problema que pode ocorrer é quando uma instituição envia várias vezes dados de uma mesma estação. Nesta caso, o migrador, a partir de informação do usuário, pode decidir ignorar ou sobrescrever os dados repetidos.

4.4.1 Tradutores de dados

A função dos tradutores de dados é receber os dados na sua forma original e convertê-los para um mesmo formato e para o esquema global apresentado na seção 4.2. Uma vez convertidos, os dados podem ser inseridos na base de dados.

Como mencionado no capítulo 3, tipicamente no Brasil as instituições que coletam dados climatológicos possuem seus próprios esquemas de modelagem de dados e formatos de armazenamento. Isto exige o desenvolvimento de um tradutor de dados para cada fonte

de dados a ser considerada. O formato original pode ser, por exemplo, arquivo ASCII, arquivo Excel ou formato relacional de um SGBD.

Na conversão dos dados para o esquema global devem ser resolvidos problemas de heterogeneidade relacionados à granularidade temporal, unidades, precisão e ocorrência de dados estimados.

Quando possível, os tradutores convertem a **granularidade** de tempo da fonte para a granularidade do esquema global (tabela 4.2, pág. 50). Por exemplo, se a fonte de dados fornece medidas de precipitação acumulada com granularidade de tempo igual a 30 minutos, o tradutor converte o conjunto de medidas para a granularidade horária ignorando algumas das medidas fornecidas. Os tradutores também convertem as medidas para as **unidades** e **precisões** utilizados no esquema global (tabela 4.1, pág. 49).

Os tradutores de dados ignoram **dados estimados** pelas fontes de dados. Assim, quando uma fonte de dados fornece um conjunto de dados no qual há medidas efetivamente coletadas e medidas estimadas a partir de medidas coletadas, as medidas estimadas não são repassadas ao migrador de dados. O objetivo desta filtragem é garantir o controle sobre a ocorrência de medidas estimadas na base de dados integrada. Não é interessante disponibilizar ao usuário medidas estimadas se não se sabe quais foram os métodos utilizados nas estimativas. As medidas estimadas, desta forma, não são armazenadas e são descartadas para sempre. Este tratamento dos dados estimados só é possível se o arquivo original da fonte de dados indicar que medidas, dentro do conjunto fornecido, são estimadas.

O formato de saída dos tradutores, introduzido no capítulo 2, é **XML**, escolhido devido a questões de interoperabilidade. Como XML é amplamente utilizado e difundido, facilita-se o intercâmbio caso se deseje fornecer os dados convertidos para outros sistemas. O Apêndice B apresenta o esquema em XML Schema a ser usado para validar os documentos XML gerados pelos tradutores. O esquema especifica como os dados e metadados fornecidos pelas fontes devem ser organizados nos documentos XML.

Opcionalmente, os tradutores podem oferecer facilidades ao usuário, como interfaces gráficas para visualização dos dados originais e dos dados convertidos, e meios de preencher manualmente lacunas de dados.

4.4.2 Migrador de dados

A principal função do migrador de dados é ler os documentos XML gerados pelos tradutores e inserir os dados convertidos na base de dados integrada. Se possível, o migrador calcula medidas diárias não fornecidas. Além disso, ele verifica algumas restrições de domínio e atualiza os dados visando consistência.

Preenchimento de valores - O migrador calcula medidas de granularidade diária quando a fonte fornece apenas medidas com granularidade horária. Isto é realizado para os casos de temperatura máxima e mínima, umidade relativa mínima, média e máxima, precipitação acumulada e radiação solar acumulada.

Verificação de consistência - O migrador analisa medidas diárias, verificando ocorrência das seguintes inconsistências:

- temperatura mínima > temperatura máxima;
- umidade relativa mínima > umidade relativa máxima; e
- medida diária fornecida incompatível com as medidas horárias fornecidas para a data em questão.

Em todos os casos o migrador ignora as medidas diárias fornecidas e, se possível, calcula novas medidas diárias utilizando as medidas horárias fornecidas.

Verificação da identificação da estação fonte - Um problema tratado pelo migrador é a questão da identificação de cada estação climatológica. Como cada fonte de dados possui uma maneira própria de identificar suas estações, não é possível utilizar na base integrada os identificadores de estações gerados pelas fontes de dados. Com isto cria-se o problema de, dada uma medida, verificar se a estação à qual ela corresponde já se encontra cadastrada na base integrada, e, em caso afirmativo, identificar a estação. A solução adotada é utilizar a localização geográfica como a maneira de identificar cada estação, sob a forma das medidas de latitude e longitude.

Outro problema está relacionado ao fornecimento, pelas fontes de dados, de medidas já fornecidas anteriormente. O migrador pode ser configurado para, nestes casos, tomar uma dentre duas ações: *ignorar* a medida caso já tenha sido fornecida uma medida para a estação e data/hora em questão, ou *substituir* a medida fornecida anteriormente pela nova medida.

4.5 Módulo de Avaliação de Qualidade

Este módulo é responsável por avaliar a qualidade dos dados fornecidos pelas fontes. A maioria das consultas sobre dados climatológicos envolve processamentos geoestatísticos que geram informações consolidadas como, por exemplo, dados espacialmente referenciados visualizados em mapas. Nestes casos, os resultados da avaliação de qualidade auxiliam o usuário a identificar os conjuntos de dados que servem melhor aos seus propósitos,

orientando a escolha dos dados a serem utilizados na consulta, e fornecem ao usuário parâmetros da credibilidade do resultado obtido.

O Módulo de Avaliação de Qualidade realiza dois tipos de avaliação, introduzidos no capítulo 2: não-dependente e dependente da tarefa. A avaliação de qualidade *não-dependente* da tarefa avalia os dados independentemente do uso pretendido. Ela é realizada após a inserção na base integrada, antes do processamento de consultas, e seus resultados são utilizados em todas as consultas que utilizam os dados avaliados. Os resultados da avaliação não-dependente da tarefa são armazenados em formato relacional no banco de dados. A avaliação *dependente* da tarefa é direcionada pelo uso que a consulta faz do dado, e portanto deve ser realizada no momento do processamento da consulta.

A próxima seção apresenta os indicadores de qualidade utilizados nas avaliações dependente e não-dependente da tarefa, e a seção seguinte descreve como os dois tipos de avaliação são aplicados na arquitetura proposta.

4.5.1 Indicadores de qualidade utilizados

A Tabela 2.1 (pág. 24) lista um conjunto de indicadores que cobre vários aspectos de avaliação de qualidade de dados. A maioria dos indicadores propostos na literatura, mesmo que com outros nomes, corresponde a um dos indicadores deste conjunto. Este conjunto serviu, portanto, como uma referência para a escolha dos indicadores a serem utilizados na arquitetura.

Após análise da relevância de cada indicador no contexto de dados climatológicos, os selecionados foram os seguintes: *ausência de erro*, *completude*, *atualidade* e *quantidade apropriada de dados*. Os dois primeiros são utilizados na avaliação não-dependente da tarefa e os dois últimos na avaliação dependente da tarefa. Como o propósito é facilitar uma avaliação objetiva da qualidade, são utilizados apenas indicadores quantitativos.

A avaliação utiliza métricas específicas para o tipo de medida em questão (precipitação, temperatura, pressão atmosférica, umidade relativa do ar ou radiação solar). As métricas da avaliação dependente da tarefa também levam em consideração a informação que se deseja obter com os dados. Todas as métricas resultam em *notas* normalizadas de 0 a 1, e devem ser determinadas por especialistas da área de climatologia. Os indicadores utilizados são descritos a seguir.

Completude

Este indicador corresponde à relação entre a quantidade de dados existente em um conjunto e a quantidade de dados esperada para o conjunto. No contexto do sistema considerado, ele é utilizado para indicar a medida da quantidade de dados faltantes em períodos de tempos das séries históricas climatológicas.

Ausência de erro

Este indicador expressa uma avaliação da medida de quanto os dados são corretos. No contexto de dados climatológicos, esta avaliação é realizada através de métodos estatísticos que comparam o valor registrado com o valor esperado, indicando possibilidade de haver dados incorretos nas séries históricas. A seção 2.5.1 apresentou algumas causas típicas deste problema para dados pluviométricos, mas as questões discutidas ali também se aplicam aos outros tipos de dados climatológicos considerados.

Atualidade

Este indicador expressa o quanto os dados são atuais para o uso pretendido. É importante verificar se os dados não são demasiadamente antigos para obter a informação desejada, uma vez que o comportamento climático de uma mesma região pode mudar com o passar do tempo. Este indicador é dependente da tarefa e portanto é calculado no momento do processamento de cada consulta.

Quantidade apropriada de dados

Este indicador expressa a medida de quanto a quantidade de dados disponível é apropriada para a tarefa em questão. Séries históricas podem não caracterizar corretamente os aspectos climáticos de uma região se não forem suficientemente extensas. Assim como atualidade, este indicador é dependente da tarefa e portanto sua avaliação é realizada no momento de cada consulta.

4.5.2 Avaliação não-dependente e dependente da tarefa

A avaliação não-dependente da tarefa é realizada sob o ponto de vista dos indicadores completude e ausência de erro. Ambos são utilizados em avaliações de *períodos de tempo* das séries históricas de cada estação climatológica. A granularidade de tempo utilizada é o *ano*.

A Figura 4.5 ilustra resultados de avaliações de qualidade dos indicadores não-dependentes da tarefa. A estação climatológica desta figura possui séries históricas de precipitação, temperatura e umidade relativa do ar. Cada ano das séries recebe uma nota para os indicadores completude e ausência de erro.

A avaliação não-dependente da tarefa deve ser realizada assim que novos anos de séries históricas são inseridos na base de dados integrada. Opcionalmente, a avaliação pode ser re-executada periodicamente, para refletir mudanças na base de dados.

Após avaliar todos os anos de uma série histórica, existe a possibilidade de aplicar uma métrica que, baseada nas notas de cada ano, produza uma nota global para toda a série.

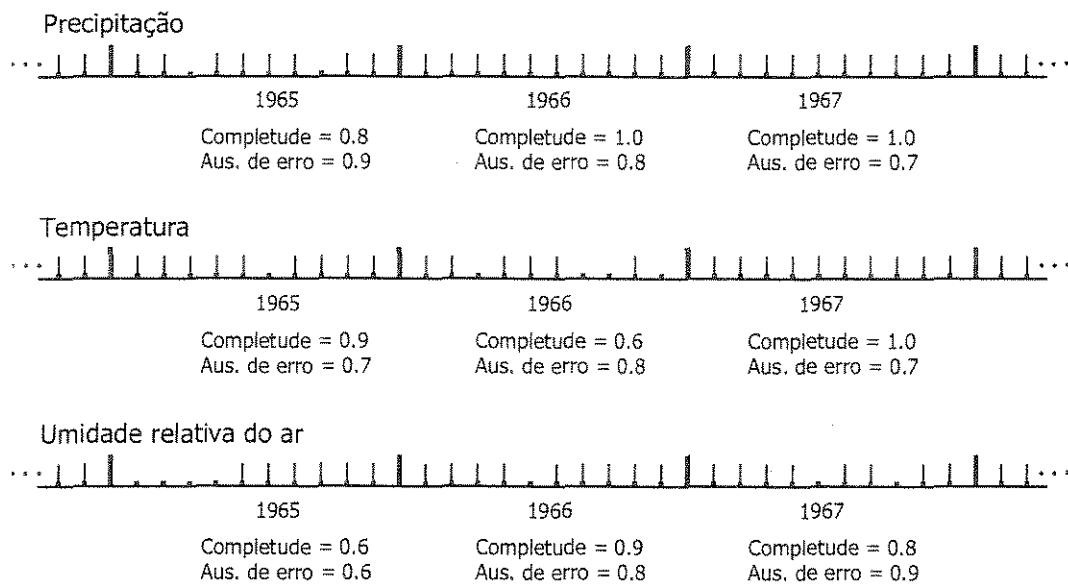


Figura 4.5: Exemplos de resultados da avaliação não-dependente da tarefa.

Esta opção não é padrão, já que as consultas sobre dados climatológicos tipicamente não utilizam as séries históricas em toda a sua extensão. Não é coerente fornecer ao usuário uma indicação da qualidade da série histórica como um todo em consultas que utilizam apenas frações da série. Da mesma maneira, descartou-se a possibilidade de aplicar uma métrica global que combinasse em uma só nota as avaliações dos quatro indicadores considerados. Esta abordagem não é coerente, pois a importância de cada indicador pode não ser a mesma para todos os usuários.

A avaliação de qualidade de dados dependente da tarefa utiliza os indicadores atualidade e quantidade apropriada de dados, sendo realizada no momento do processamento de cada consulta. Na solicitação de uma consulta, portanto, o Módulo de Avaliação de Qualidade interage com o Módulo de Processamento de Consultas como descreve a próxima seção.

4.6 Módulo de Processamento de Consultas

A função básica deste módulo é permitir a realização de consultas sobre a base de dados integrada. Estão previstas consultas tradicionais (alfanuméricas) e consultas espaciais. Exemplos de produtos resultantes são distribuição espacial de valores para determinados períodos, identificação de extremos, cálculo de estatísticas e variações espaço-temporais de aspectos climáticos.

Após estudos das necessidades típicas de usuários da área de climatologia, alguns requisitos do módulo foram levantados e são discutidos a seguir. Considera-se que a interface utilizada para a realização de consultas pertence ao escopo do Módulo de Processamento de Consultas.

4.6.1 Funções básicas

Disponibilização de dados - O módulo deve fornecer ao usuário maneiras de determinar que conjuntos de dados devem ser utilizados na consulta. O usuário deve ser capaz de escolher os dados utilizando o banco de metadados e os resultados da avaliação de qualidade não-dependente da tarefa.

Estatísticas básicas - O módulo deve ser capaz de processar consultas simples, como médias, extremos, totais e outras estatísticas para determinadas regiões e períodos de tempo.

Visualização de resultados - Os resultados podem ser apresentados sob forma textual (valores, tabelas), gráfica (histogramas, curvas de variação de valores, mapas) ou via geração de arquivos (planilhas, mapas digitais). Em consultas que envolvam cálculos de distribuição espacial de medidas climatológicas, o Módulo de Processamento de Consultas deve interagir com algum SIG. Na execução da consulta, o módulo pode solicitar ao SIG a geração de um mapa estático, na forma de um arquivo de imagem, e exibir o mapa utilizando a interface do próprio sistema. Outra possibilidade é o sistema gerar um mapa no formato do SIG, permitindo ao usuário interagir com o mapa posteriormente utilizando a interface do SIG.

Caracterizações climáticas - Outra funcionalidade a ser disponibilizada pelo Módulo de Processamento de Consultas é o cálculo de caracterizações climáticas de estações e regiões a partir das séries históricas integradas, orientando a programação agrícola de cada região. Estas caracterizações devem auxiliar alertas de ocorrências de eventos climáticos prejudiciais à agricultura, como chuvas extremas, veranicos e geadas, possibilitando a tomada das medidas preventivas necessárias.

4.6.2 Resultados de consultas

Na realização de uma consulta, deve-se informar ao usuário, além dos dados resultantes da consulta, o resultado das avaliações de qualidade não-dependente e dependente da tarefa. Para tanto, o Módulo de Processamento de Consultas interage com o Módulo de Avaliação de Qualidade.

A Figura 4.6 exibe de maneira esquemática como ocorre esta interação. Após realizar a consulta propriamente dita, o Módulo de Processamento de Consultas informa ao Módulo de Avaliação de Qualidade os dados utilizados e o tipo de informação solicitada. O Módulo de Avaliação de Qualidade realiza a avaliação de qualidade e informa os resultados ao Módulo de Processamento de Consultas, que os repassa à interface, juntamente com o resultado da consulta.

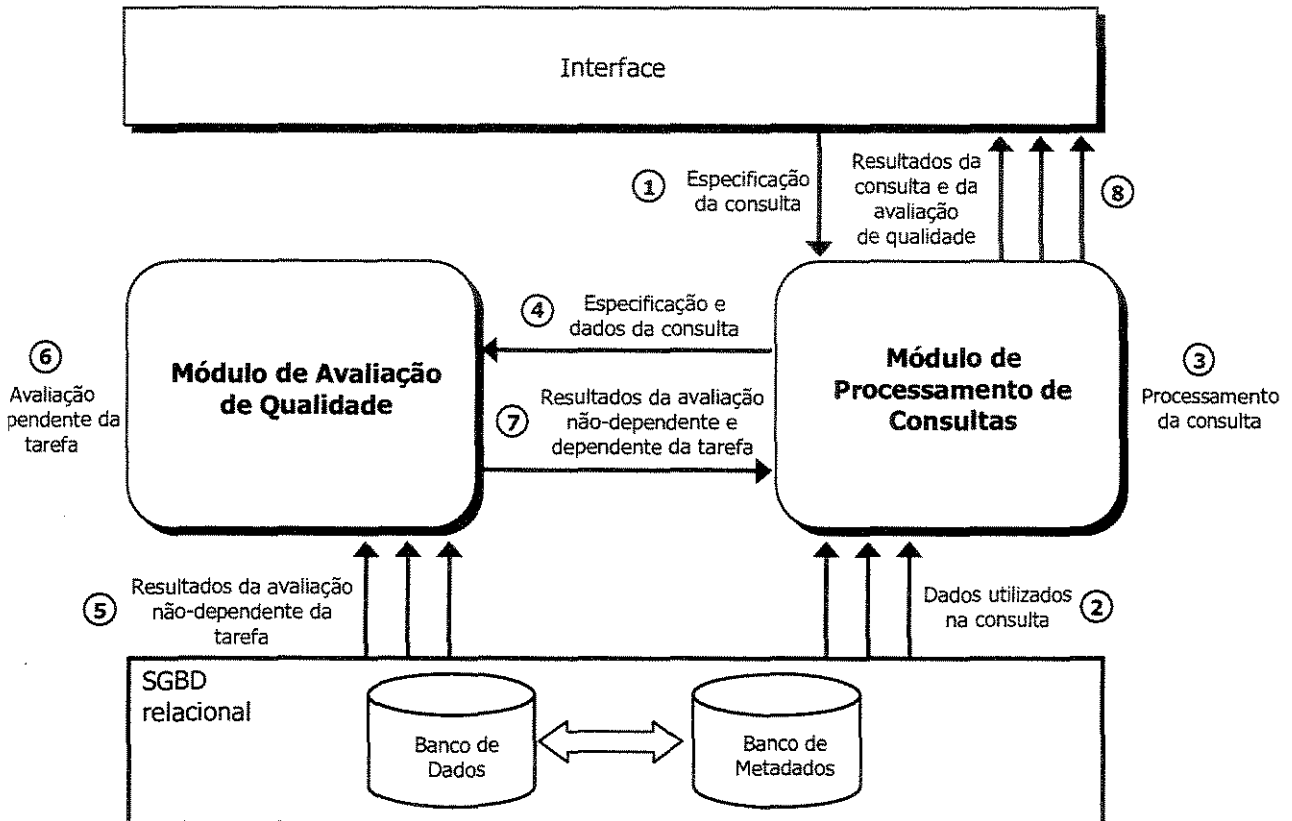


Figura 4.6: Interação entre os módulos de Avaliação de Qualidade e de Processamento de Consultas.

Veamos um exemplo de consulta, na qual o usuário solicita ao sistema a geração de um gráfico das *médias de totais mensais pluviométricos* da estação Bragança Paulista. Inicialmente o usuário especifica a estação e a faixa de anos a serem considerados na geração do gráfico. Os passos do processamento da consulta são os seguintes:

1. A interface do sistema repassa ao Módulo de Processamento de Consultas a especificação da consulta.

2. O Módulo de Processamento de Consultas acessa o banco de dados integrado e realiza a consulta propriamente dita.
3. O Módulo de Processamento de Consultas realiza a consulta propriamente dita.
4. O Módulo de Processamento de Consultas informa ao Módulo de Avaliação de Qualidade os dados utilizados na consulta, e que a informação desejada são *médias de totais mensais pluviométricos*.
5. O Módulo de Avaliação de Qualidade acessa a base de dados para obter os resultados da avaliação de qualidade não-dependente da tarefa, e consolida as notas dos indicadores completude e ausência de erro atribuídas anteriormente aos anos utilizados na consulta, produzindo uma nota para cada indicador (notar que neste exemplo trata-se de apenas uma estação).
6. O Módulo de Avaliação de Qualidade realiza a avaliação dependente da tarefa. De acordo com a métrica elaborada para o produto *médias de totais mensais pluviométricos*, ele avalia os anos utilizados na consulta, produzindo uma nota para os indicadores atualidade e quantidade apropriada de dados.
7. O Módulo de Avaliação de Qualidade repassa ao Módulo de Processamento de Consultas os resultados da avaliação de qualidade.
8. O Módulo de Processamento de Consultas repassa os resultados da consulta e da avaliação de qualidade à interface, que apresenta o resultado da consulta graficamente.

Se, por outro lado, em vez de um gráfico o usuário solicita um *mapa de distribuição de totais pluviométricos mensais* no estado de São Paulo, os passos 3 e 8 são modificados da seguinte forma:

3. O Módulo de Processamento de Consultas extrai da base integrada os conjuntos de dados a serem utilizados, incluindo o georeferenciamento das estações origem.
 - 3.1. O conjunto de dados georeferenciado é repassado a um SIG, que produz o mapa desejado.
8. O mapa e os resultados da avaliação de qualidade são repassados à interface para visualização.

Este exemplo de consulta, diferentemente do anterior, utiliza várias estações origem. Neste caso, o Módulo de Avaliação de Qualidade calcula uma nota para cada indicador de qualidade para *cada estação* utilizada na consulta.

4.7 Resumo

Este capítulo apresentou a solução proposta para o problema de integração de dados climatológicos sob a forma de uma arquitetura para um sistema centralizado. Foram descritos alguns aspectos da arquitetura, incluindo a modularização do sistema; a modelagem do banco de dados integrado; a especificação do banco de metadados e como este se relaciona com o banco de dados; a metodologia de integração do Módulo de Integração; a metodologia de avaliação de qualidade do Módulo de Avaliação de Qualidade e as funções básicas do Módulo de Processamento de Consultas.

Capítulo 5

Aspectos de implementação

A arquitetura proposta neste trabalho foi validada através da implementação parcial de um sistema de integração de dados pluviométricos. A implementação fez parte de um projeto de monitoramento agrometeorológico desenvolvido na Empresa Brasileira de Pesquisa em Agropecuária (EMBRAPA). Neste projeto, o sistema de integração de dados pluviométricos faz parte de um sistema chamado **Agritempo**. O back-end do sistema foi implementado em PL/SQL, e o SGBD utilizado foi OracleTM [22] versão 8.1.6. Este capítulo apresenta os principais aspectos da implementação.

A seção 5.1 descreve o banco de dados do sistema. A seção 5.2 apresenta alguns aspectos da implementação do Módulo de Integração de Dados. Em seguida, a seção 5.3 descreve a implementação do Módulo de Avaliação de Qualidade e a seção 5.4 apresenta algumas das dificuldades ocorridas na implementação.

É importante notar que a implementação foi guiada principalmente pelos requisitos de usuários levantados na EMBRAPA.

5.1 Banco de Dados

Para atender aos requisitos de usuário, incluindo os requisitos de avaliação de qualidade, o sistema deve manter informações a respeito de:

- instituições fornecedoras de dados;
- estados;
- municípios;
- estações pluviométricas;
- medidas pluviométricas diárias;

- meses de séries históricas;
- anos de séries históricas; e
- zonas homogêneas.

A Figura 5.1 apresenta o diagrama entidade-relacionamento do banco de dados implementado, que estende a Figura 4.2 com as entidades Ano de série, Zona Homogênea, Total mensal e Estado. Neste diagrama atributos compostos têm seus componentes especificados entre parênteses e atributos multivalorados são indicados por um asterisco. A seguir as entidades são brevemente descritas.

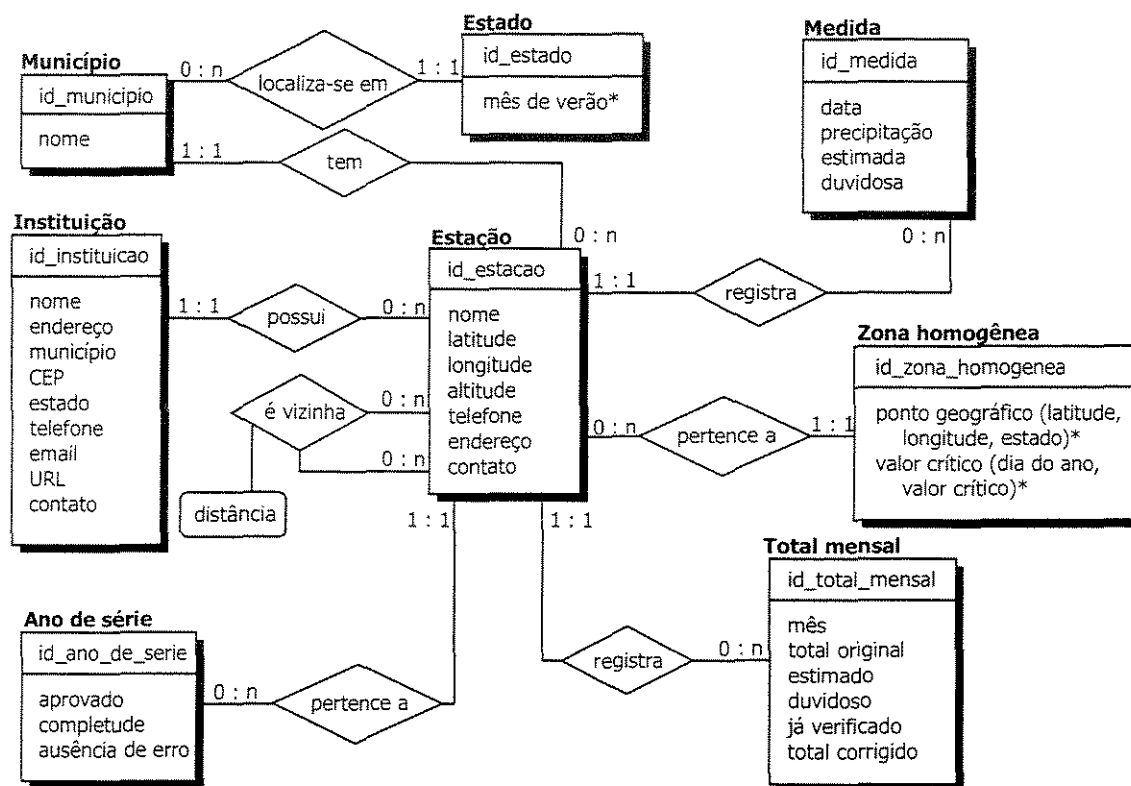


Figura 5.1: Diagrama entidade-relacionamento.

Instituição - Entidade que representa as instituições que fornecem dados ao sistema. Exemplos de instituições incluem o Departamento de Águas e Energia Elétrica do Estado de São Paulo, o Ministério da Agricultura e a Agência Nacional de Águas. Cada instituição pode ter várias estações pluviométricas.

Município - Representa os municípios onde se localizam as estações pluviométricas. Cada município pertence a um estado e pode possuir várias estações.

Estado - Representa os estados brasileiros. Cada estado possui um ou mais meses considerados como período de verão. Neste contexto, considera-se verão o(s) período(s) onde há alta frequência de ocorrência de chuvas. Os períodos de verão são utilizados na avaliação de qualidade das séries históricas pluviométricas.

Estação - Representa as estações que registram medidas pluviométricas. Cada estação pertence a uma instituição, localiza-se em um município e em uma zona homogênea, pode possuir várias medidas pluviométricas diárias, anos de séries e totais mensais, e pode ser vizinha de outras estações. Duas estações são consideradas vizinhas se a distância entre elas é menor que 37 quilômetros, e se ambas se encontram na mesma faixa de altitude, considerando as seguintes faixas: 0 a 300 metros de altitude, 300 a 850 metros, e acima de 850 metros.

Medida - Representa as medidas pluviométricas das estações. Cada medida corresponde ao total pluviométrico ocorrido na estação em uma *data* específica. O atributo *estimada* indica se a medida foi estimada pelo sistema, e o atributo *duvidosa*, se a medida foi considerada duvidosa na etapa de avaliação de qualidade.

Total mensal - Representa os totais pluviométricos de meses completos da série histórica de uma estação. Uma instância desta entidade pode corresponder, por exemplo, ao total pluviométrico do mês de janeiro de 1978 na estação São Bernardo do Campo. Os atributos *estimado* e *duvidoso* indicam se há medidas diárias estimadas ou duvidosas no mês em questão. O atributo *já verificado* indica se o mês já foi verificado na etapa de avaliação de qualidade, a qual pode registrar, se for o caso, um *total corrigido* para o mês, mantendo o total pluviométrico original no atributo *total original*.

Ano de série - Representa os anos da série histórica de uma estação. Ao contrário da entidade **Total mensal**, não é necessário que o ano esteja completo para ser registrado. O atributo *aprovado* indica se o ano foi aprovado na avaliação de qualidade. Nesta etapa o ano recebe um valor para os indicadores *completude* e *ausência de erro*.

Zona homogênea - Representa regiões geográficas definidas de acordo com o comportamento pluviométrico, de maneira que estações que se encontrem em uma mesma zona homogênea apresentem comportamento semelhante. A região de cada zona é definida por *pontos geográficos*, e cada zona possui um *valor crítico* a 3% para cada dia do ano. A

probabilidade de ocorrer uma quantidade de precipitação maior que a indicada pelo valor crítico, no dia do ano e zona em questão, é inferior a 3%. Estes valores não são calculados pelo sistema, mas definidos pelo Grupo de Estatística do Projeto de Zoneamento e Monitoramento Agrícolas do Ministério da Agricultura, e são utilizados na avaliação de medidas pluviométricas diárias.

A Figura 5.2 exibe as relações correspondentes ao diagrama entidade-relacionamento da Figura 5.1. As relações que pertencem exclusivamente ao sistema de integração de medidas pluviométricas possuem prefixo *sh*, correspondente a “séries históricas”. As demais relações são utilizadas em outros módulos do sistema Agritempo.

As relações *Instituicao*, *Municipio* e *Estacao* implementam, respectivamente, as entidades **Instituição**, **Município** e **Estação**. Elas possuem atributos adicionais aos especificados para estas entidades, necessários a outros módulos do sistema Agritempo.

A relação *sh_estacao_vizinha* implementa o auto-relacionamento “é vizinha” da entidade **Estação**. Ela é mantida atualizada através de gatilhos sobre a relação *Estacoes*.

A relação *sh_periodo_verao* implementa o atributo multivalorado *mês de verão* da entidade **Estado**. Cada tupla desta relação corresponde ao registro de um mês considerado verão para um determinado estado. Como a entidade **Estado** não possui outros atributos, não foi necessário criar uma relação para implementá-la.

As relações *sh_medida*, *sh_total_mensal* e *sh_ano_de_serie* implementam, respectivamente, as entidades **Medida**, **Total Mensal** e **Ano de série**. As duas últimas relações são mantidas atualizadas através de gatilhos sobre a relação *sh_medida*. O atributo *data_armazenamento*, da relação *sh_medida*, corresponde ao tempo de transação, e indica a data em que a medida foi inserida na base de dados, sendo atualizada sempre que a medida tem seu valor alterado.

A relação *sh_zona_homogenea* implementa a entidade **Zona homogênea**. A relação *sh_composicao_zona_homogenea* implementa o atributo multi-valorado *ponto geográfico* desta entidade, e a relação *sh_valor_critico*, o atributo multi-valorado *valor crítico*.

A relação auxiliar *sh_verificacao_zona* foi criada para agilizar a inserção de estações. Quando uma nova estação é inserida na base de dados, é necessário identificar a que zona homogênea ela pertence, baseando-se nos pontos geográficos que compõem cada zona. Isto é realizado através do método *vizinho mais próximo*: o sistema calcula a distância entre a estação sendo inserida e os pontos que compõem as zonas homogêneas. A estação é associada à zona que possui o ponto mais próximo. Para agilizar o processo, são calculadas somente as distâncias para os pontos que se encontrarem em estados vizinhos ao estado da estação sendo inserida. A relação *sh_verificacao_zona* indica, para cada estado, os estados cujos pontos devem ter sua distância verificada.

A relação *sh_configuracao* também é uma relação auxiliar, que possui uma única

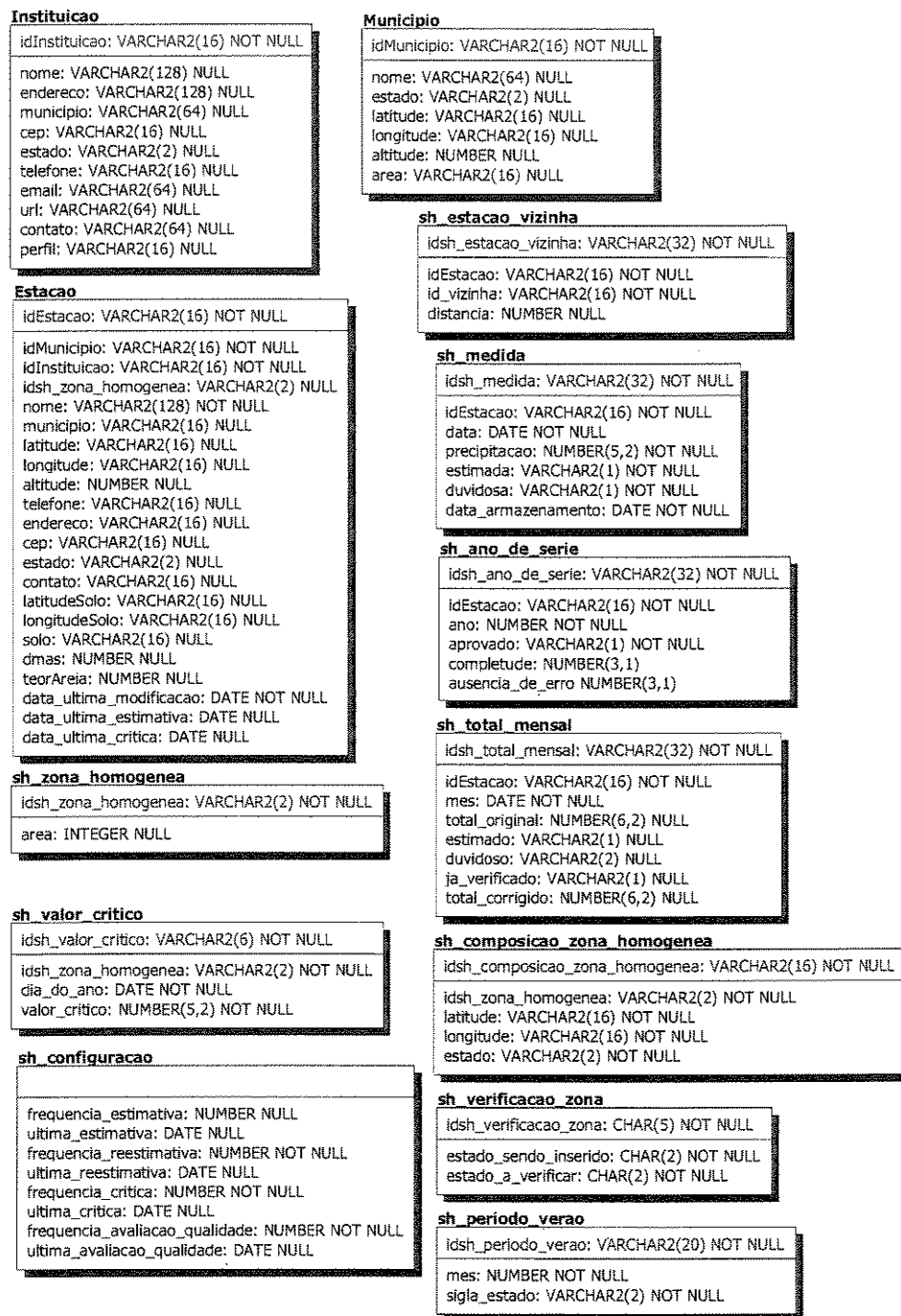


Figura 5.2: Esquema resultante do diagrama entidade-relacionamento.

tupla onde são registrados alguns parâmetros de configuração do sistema. Nela é armazenada a frequência, em dias, com a qual devem ser realizados os processos de estimativa de medidas, reestimativa daquelas já estimadas e avaliação de qualidade não-dependente da tarefa, bem como a data da última realização destes processos.

5.2 Módulo de Integração

Os tradutores foram customizados e implementados para cada instituição que fornece dados ao sistema. Alguns foram desenvolvidos em JavaTM[49], e outros em Visual Basic for ApplicationsTM[21], estes dentro do aplicativo Microsoft Excel. Os tradutores foram implementados por Luciana Romani, Arnaldo Montagner e Edgar Santos, da equipe do projeto de agrometeorologia da EMBRAPA.

Os usuários dos tradutores são colaboradores das instituições fornecedoras de dados. Os tradutores oferecem algumas facilidades ao usuário, como interfaces gráficas para preenchimento manual de lacunas e capacidade de converter em uma só execução vários arquivos de entrada em um arquivo de saída.

A Figura 5.3 exibe um trecho de arquivo de dados no formato original da instituição. Neste exemplo, o formato é ASCII, e cada linha do arquivo contém 12 medidas pluviométricas diárias de uma estação. Medidas assinaladas com um asterisco foram estimadas pela instituição, sendo portanto ignoradas pelo tradutor.

```
E2-098 88 11 .8 .0 .1 .0 .2 .1 .0 .0 .0 .3 .0
E2-098 88 12 .0 .0 .0 24.7 .1 .0 .1 .0 .0 .0 .0 .1
E2-098 88 13 .0 .0 .0 .0 .0 .5 .0 .0 .0 .1 .0 .0
E2-098 88 14 .0 .1 .0 .1 .0 .0 .0 .0 .0 .0 .2 .0
E2-098 88 15 .0 .8 .0 .0 .2 .0 .1 .0 13.5 .0 1.5 .0
E2-098 88 16 .2 .2 .0 6.2 .8 .0 .0 .0 9.8 .1 .3 .0
E2-098 88 17 1.2 13.9 .0 .0 .1 .0 .0 .0 3.2 .3 .0 .0
E2-098 88 18 .0 16.1 .0 .1 17.2 .0 .0 .0 10.2 15.9 .1 .8
E2-098 88 19 .0 2.3 .1 .6 12.8 .0 .0 .0 .9 12.8 .4 .0
E2-098 88 20 .0 25.9 .2 12.1 11.6 .0 .1 .0 .1 .0 .1 .3
E2-098 88 21 .0 13.9 21.6 9.8 .1 .0 .0 .0 .0 .1 .8 .7
E2-098 88 22 .0 1.3 8.6 .0 1.7 .6 .0 .0 .1 .8 .1 .2
E2-098 88 23 14.6 .0 .0 .0 9.2 2.5 .0 .0 .6 .0* 6.3 .4
E2-098 88 24 79.1 .0 .0 .0 18.2 .0 .0 .0 .0 .0* .0 .1
E2-098 88 25 .0 .0 .0 .1 .5 .0 .0 .0 .8 .0* .0 .8
E2-098 88 26 7.3 18.6 .2 .2 .3 .0 .0 .0 .2 6.1* .2 1.6
E2-098 88 27 1.3 .0 .0 .0 .1 .1 .0 .0 .1 .1* 13.8 .6
E2-098 88 28 .0 .0 .0 .0 13.4 .1 .0 .0 .0 .0 60.9 .1
E2-098 88 29 .0 .0 .0 .3 .2 .0 .0 .1 .0 1.2 .2 8.3
```

Figura 5.3: Exemplo de arquivo de dados no formato original da instituição.

A Figura 5.4 exibe um trecho de um arquivo de saída dos tradutores. Na proposta,

```
01011912;SAO LUIZ DO CURU;CE;0
02011912;SAO LUIZ DO CURU;CE;0
03011912;SAO LUIZ DO CURU;CE;5
04011912;SAO LUIZ DO CURU;CE;0
05011912;SAO LUIZ DO CURU;CE;4
06011912;SAO LUIZ DO CURU;CE;0
07011912;SAO LUIZ DO CURU;CE;12
08011912;SAO LUIZ DO CURU;CE;0
09011912;SAO LUIZ DO CURU;CE;0
10011912;SAO LUIZ DO CURU;CE;0.8
11011912;SAO LUIZ DO CURU;CE;0
12011912;SAO LUIZ DO CURU;CE;0
13011912;SAO LUIZ DO CURU;CE;0
14011912;SAO LUIZ DO CURU;CE;0
15011912;SAO LUIZ DO CURU;CE;0
16011912;SAO LUIZ DO CURU;CE;70
17011912;SAO LUIZ DO CURU;CE;0
18011912;SAO LUIZ DO CURU;CE;8
19011912;SAO LUIZ DO CURU;CE;39.2
20011912;SAO LUIZ DO CURU;CE;0
21011912;SAO LUIZ DO CURU;CE;0
```

Figura 5.4: Exemplo de arquivo de saída dos tradutores.

o formato dos arquivos de saída é XML, mas na Embrapa optou-se por implementar utilizando formato ASCII. Cada linha do arquivo contém uma data, o nome e o estado de uma estação e a quantidade total de chuva correspondente. Na relação *Estacao*, o par formado pelo nome da estação e seu estado é único.

O **migrador** foi implementado em JavaTM. Ele lê os arquivos gerados pelos tradutores, acessa a base de dados OracleTM e insere as medidas pluviométricas. Ele também realiza as ações descritas no capítulo 4: identificação, e, se for o caso, cadastro da estação; validações básicas sobre as medidas e tratamento de medidas já fornecidas anteriormente.

O migrador é executado automaticamente todos os dias, em uma hora configurável. Ele identifica os novos arquivos de cada instituição, recebidos via FTP, processa-os, e gera arquivos de *log* com eventuais problemas ocorridos durante o processo de inserção das medidas.

5.3 Módulo de Avaliação de Qualidade

A implementação da avaliação de qualidade foi dividida em duas partes. A primeira consiste em uma etapa de verificação da consistência das séries históricas. A segunda consiste na avaliação de qualidade não-dependente da tarefa, como descrito no capítulo 4. As duas próximas seções detalham estas implementações.

5.3.1 Consistência de séries históricas

O objetivo desta etapa é preparar as séries históricas para a avaliação de qualidade não dependente da tarefa e para o processamento de consultas. Ela consiste em quatro partes: análise de medidas diárias, estimativa de medidas diárias faltantes, crítica de séries e anos de séries e análise de totais mensais. Todas as verificações foram implementadas via *stored procedures* em PL/SQL. Desta forma, elas são executadas no servidor do banco de dados.

Os critérios aplicados nas verificações foram definidos por especialistas da área de climatologia, na EMBRAPA.

Análise de medidas diárias

A verificação de medidas diárias identifica e assinala medidas diárias consideradas “duvidosas”. A quantidade de medidas duvidosas nas séries históricas é utilizada na avaliação de qualidade não-dependente da tarefa, como descrito adiante. Além disto, quando o usuário realiza uma consulta, o sistema pode lhe informar a porcentagem de medidas duvidosas utilizadas, fornecendo mais parâmetros para tomadas de decisão.

Uma medida é considerada duvidosa quando seu valor é maior que 150mm, ou quando seu valor é maior que o valor crítico da zona homogênea em que se encontra a estação correspondente, no dia do ano em questão.

A análise de medidas diárias é executada via gatilho sobre a relação `sh_medida`. No momento da inserção ou alteração de uma medida, seu valor é verificado segundo os critérios descritos, marcando-se as medidas duvidosas.

Estimativa de medidas diárias faltantes

O método escolhido para implementar a estimativa de medidas diárias faltantes foi uma combinação de alguns dos métodos descritos na seção 2.5.1. Toma-se como base a precipitação, da data em questão, das n estações vizinhas à estação x que apresenta a falta do dado. A precipitação faltante $p_x(t_f)$ é estimada pela média aritmética ponderada das medidas das estações vizinhas, onde o peso aplicado é o inverso do quadrado da distância L entre a estação vizinha e a estação x . A fórmula resultante é a seguinte:

$$p_x(t_f) = \frac{\sum_{i=1}^n [p_i(t_f)]}{\sum_{i=1}^n (\frac{1}{L_i^2})} \quad (5.1)$$

Apesar de se registrar em `sh_estacao_vizinha` as estações que se encontrem a menos de 37 quilômetros da estação x , no momento da estimativa inicialmente considera-se apenas as estações vizinhas que se encontrem a menos de 22 quilômetros da estação x . Se há pelo menos três vizinhas nesta situação, somente as medidas dentro deste raio são consideradas. Se não há, considera-se então todas as estações vizinhas. Esta restrição

contribui para a acurácia das estimativas, uma vez que estações mais próximas possuem comportamento pluviométrico mais parecido.

As medidas estimadas são armazenadas na relação `sh_medida`, juntamente com as medidas efetivamente coletadas. Toda medida estimada é marcada como tal, para que no momento da especificação de consultas o usuário possa optar por utilizar ou não medidas estimadas.

A estimativa de medidas pode ser executada manualmente, especificando-se uma estação e período de tempo. Neste caso o sistema tenta estimar todas as medidas faltantes da estação em questão no período especificado. A Figura 5.5 exibe a tela na qual o curador do banco de dados especifica a estação cujas medidas devem ser estimadas. A estimativa também é realizada automaticamente, com frequência configurável, para refletir mudanças na base de dados. A cada n dias é executado um processo que tenta estimar as medidas ainda faltantes para cada estação cujas vizinhas possuem ao menos uma nova medida registrada desde a última tentativa de estimativa realizada para aquela estação. O parâmetro n é configurado no atributo `sh_configuracao.frequencia_estimativa`, e as datas da última modificação de medidas e da última tentativa de estimativa de cada estação são armazenadas, respectivamente, nos atributos `data_ultima_modificacao` e `data_ultima_estimativa` da relação `estacao`.

Também são executadas automaticamente, com frequência configurável, reestimativas de medidas já estimadas, utilizando dados de estações vizinhas inseridos ou atualizados após a última estimativa. O processo de reestimativa recalcula estimativas quando há modificação nos dados das vizinhas dentro de um período de tempo. Este processo, ao refletir mudanças ocorridas nas estações vizinhas, contribui para refinar as estimativas. A frequência com a qual a reestimativa deve ser executada é configurada no atributo `sh_configuracao.frequencia_reestimativa`.

Crítica de séries e anos

O objetivo da verificação de séries e anos é identificar quais séries históricas e anos de séries são adequados para serem submetidos à avaliação de qualidade, e, posteriormente, ao processamento de consultas.

Os critérios aplicados são aqueles descritos na seção 3.2.2. Quando a série como um todo é reprovada na crítica, todos os seus anos são marcados como reprovados na relação `sh_ano_de_serie`. Quando apenas alguns de seus anos são reprovados, apenas estes são marcados como reprovados naquela relação. Os anos de séries marcados como reprovados não são submetidos à avaliação de qualidade e tampouco ao processamento de consultas.

A crítica de séries históricas pode ser solicitada manualmente, especificando-se a estação pluviométrica cuja série deseja-se avaliar. A crítica também é realizada automaticamente, com frequência configurável, para refletir mudanças na base de dados. A

Agritempo

Sistema de Monitoramento Agroclimológico

Séries Históricas de Chuvas
Estimativa de medidas diárias faltantes - IAC

Voltar

Estações disponíveis

Estação	UF	Latitude	Longitude	Data início	Data fim	Nº medidas faltantes	Nº medidas estimadas	Nº estações vizinhas	
BRASILIA(Brasília)	GO	15o46'S	47o55'W	15/07/1996	20/05/2001	150	0	2	Estimar
CATALAO(Catalão)	GO	18o06'S	47o34'W	15/07/1996	20/05/2001	25	0	0	
FORMOSA(Formosa)	GO	15o31'S	47o19'W	14/09/1995	20/05/2001	30	0	0	
GOIANESIA(Goianesia)	GO	15o18'S	49o12'W	15/07/1996	20/05/2001	70	4	3	Estimar
GOIANIA(Gorânia)	GO	16o36'S	49o18'W	15/07/1996	20/05/2001	284	0	0	
GOIAS(Goiás)	GO	15o48'S	50o18'W	15/07/1996	20/05/2001	0	0	1	Estimar
IPAMERI(Ipameri)	GO	17o24'S	48o00'W	15/07/1996	20/05/2001	15	5	0	
ITUMBARA(Itumbara)	GO	18o24'S	49o18'W	15/07/1996	04/03/1997	78	0	4	Estimar
JATAI(Jatai)	GO	17o42'S	51o42'W	15/07/1996	20/05/2001	115	0	5	Estimar
PIRENOPOLIS(Pirenópolis)	GO	15o42'S	49o00'W	15/07/1996	20/05/2001	20	3	0	
POSSE(Posse)	GO	14o12'S	46o30'W	15/07/1996	20/05/2001	45	0	1	Estimar
RIOVERDE(Rio Verde)	GO	17o36'S	51o06'W	15/07/1996	20/05/2001	12	0	0	

Figura 5.5: Tela de estimativa de medidas faltantes.

cada p dias é executado um processo que critica todas as séries cujo conjunto de medidas tenha sofrido alguma modificação desde a sua última crítica. O parâmetro p é configurado no atributo `sh_configuracao.frequencia_critica`, e as datas da última crítica e da última modificação de medidas de cada estação são armazenadas, respectivamente, nos atributos `data_ultima_critica` e `data_ultima_modificacao` da relação `estacao`.

Análise de totais mensais

A verificação de totais mensais tem como objetivo identificar totais pluviométricos mensais considerados “duvidosos”. Ela se justifica pelo fato de diversas consultas típicas da área de pluviometria se basearem em totais mensais.

Ela é realizada utilizando o método estatístico *vetor regional* [64]. Neste método analisa-se um período de n meses das séries de um conjunto de estações pluviométricas que pertençam a uma mesma região. Mediante processo iterativo, tenta-se determinar

um vetor de n valores que caracteriza o comportamento da chuva na região nos n meses em questão; maiores detalhes podem ser obtidos em [64].

O resultado é um novo conjunto de totais mensais estimados, que podem ser comparados aos totais mensais originais. Caso desejado, os totais mensais estimados são armazenados na base de dados, mantendo-se os totais originais. Ao especificar uma consulta, o usuário tem a opção de escolher entre totais mensais originais ou corrigidos.

5.3.2 Avaliação não-dependente da tarefa

A avaliação não-dependente da tarefa atribui, para cada ano de série histórica, uma nota aos indicadores completude e ausência de erro.

A métrica utilizada para o indicador *completude* toma como base o número de medidas faltantes no ano sendo avaliado. Quanto maior este número, menor a nota associada. A nota é determinada como a razão entre o número de medidas existentes para o ano e o número máximo de medidas possível (365), resultando em um número entre 0 e 1 como proposto no capítulo 4. Assim, um ano que apresenta 15% de medidas faltantes recebe nota 0.85.

A métrica utilizada para o indicador *ausência de erro* segue o mesmo princípio, mas levando em consideração o número de medidas duvidosas do ano sendo avaliado. Quanto maior o número de medidas duvidosas no ano, menor a nota associada. A nota é determinada como a razão entre o número de medidas não-duvidosas existentes no ano e o número total de medidas possível (365). Por exemplo, um ano que possui 7% de medidas duvidosas recebe nota 0.93.

Descartou-se a possibilidade de definir uma métrica que atribuísse uma nota para cada faixa de porcentagem, como, por exemplo, atribuir à completude uma nota 0.9 para anos que possuíssem de 10% a 20% de medidas faltantes. A expressão da nota de cada ano da maneira implementada permite a cada usuário aplicar seus próprios critérios de avaliação e aceitação, como discutido no capítulo 3.

A avaliação de qualidade não-dependente da tarefa pode ser executada manualmente, especificando-se a estação desejada. Ela também é executada automaticamente, com frequência configurável, para refletir mudanças na base de dados. Ela é configurada como os processos de crítica e estimativa, descritos anteriormente, no atributo `sh_configuracao.frequencia_avaliacao_qualidade`.

5.3.3 Visões

Foram implementadas algumas *visões* que sumarizam resultados do Módulo de Avaliação de Qualidade, tendo duas funções principais: controlar e facilitar o acesso aos dados integrados. As visões e seus atributos são exibidas na Figura 5.6.

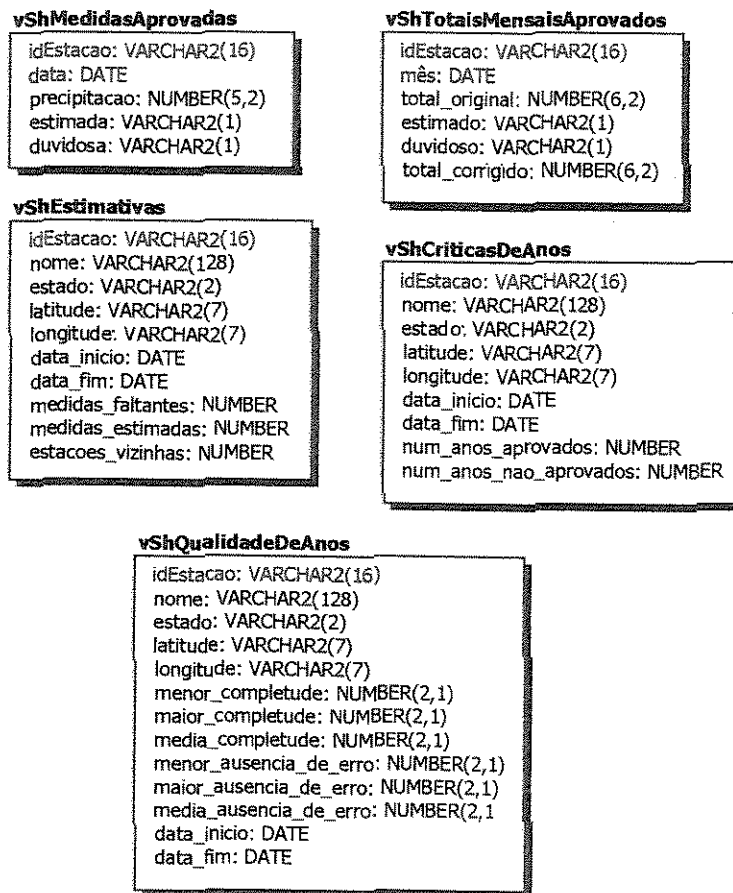


Figura 5.6: Visões implementadas.

As visões `vShMedidasAprovadas` e `vShTotaisMensaisAprovados` são utilizadas para determinar que dados podem ser acessados pelo Módulo de Processamento de Consultas. Ambas se baseiam nos resultados da crítica de séries e anos descrita na seção 5.3.1. A primeira disponibiliza todas as medidas diárias pertencentes a anos aprovados na etapa de crítica. A segunda disponibiliza todos os totais mensais pertencentes a anos aprovados na crítica. O Módulo de Processamento de Consultas, ao utilizar estas visões para acessar os dados, assegura que apenas dados aprovados na crítica são disponibilizados ao usuário.

A visão `vShQualidadeDeAnos` é utilizada para informar dados sobre os resultados da avaliação de qualidade não-dependente da tarefa, auxiliando o usuário na escolha dos dados a serem utilizados em cada consulta. Para cada estação, ela informa dados como a média, a menor e a maior das notas de completude atribuídas aos anos da série, a média, a menor e a maior das notas de ausência de erro, e os anos inicial e final da série histórica. São considerados apenas os anos aprovados na etapa de crítica, uma vez que apenas estes podem ser utilizados nas consultas.

A visão `vShCriticasDeAnos` fornece estatísticas sobre os resultados das críticas de anos de séries. Ela é utilizada pelo curador do banco de dados para avaliar o estado da base integrada. Ela informa, para cada estação, dados como datas de início e fim da série, número de anos aprovados, número de anos não-aprovados e número total de anos da série. Séries históricas com número excessivo de anos não-aprovados podem ser examinadas pelo curador, que pode tomar providências como consultar a instituição fonte de dados sobre a existência de mais medidas para a estação.

A visão `vShEstimativas` sumariza dados a respeito de estimativas de medidas faltantes. Ela também é utilizada pelo curador para avaliar o estado do banco de dados, e para direcionar a execução manual do processo de estimativas de medidas faltantes. Para cada estação, esta visão informa dados como datas de início e fim da série histórica, número de medidas já estimadas para a estação, número de medidas ainda faltantes e número de estações vizinhas existentes para a estação.

Neste trabalho o Módulo de Processamento de Consultas foi levado apenas até a fase de discussão e análise. A implementação deste módulo deve seguir a Especificação de Requisitos do capítulo 3 e as discussões do capítulo 4.

5.4 Dificuldades da implementação

A principal dificuldade de implementação ocorreu devido ao grande volume de dados manipulado. O fato de se lidar com inserções e atualizações de mais de 6 milhões de medidas gerou altos custos computacionais, exigindo grande capacidade de processamento

do servidor do banco de dados e limitando a implementação. As criações de visões foram dificultadas, impedindo a execução *on-line* dos processos do Módulo de Avaliação de Qualidade. Foi necessário realizar *tuning* sobre as visões criadas, e implementar as etapas de crítica de anos, estimativa de medidas faltantes e avaliação de qualidade não-dependente da tarefa como processos *batch*, em vez de implementá-los através de gatilhos.

Uma outra dificuldade, não esperada a princípio, ocorreu na implementação dos gatilhos sobre a relação `sh_medida`, responsáveis, entre outros, por manter atualizadas as relações `sh_ano_de_serie` e `sh_total_mensal`. Somente os procedimentos destas atualizações exigiram 16 horas de testes.

5.5 Resumo

Este capítulo apresentou os principais aspectos da implementação para o caso real da EMBRAPA: como o esquema proposto para o banco de dados foi adaptado para o caso real; a implementação dos tradutores de dados; o desenvolvimento do Módulo de Avaliação de qualidade, incluindo particularidades dos métodos de consistência de séries históricas e de sua automação. O capítulo também apresentou as principais dificuldades ocorridas na implementação.

Capítulo 6

Conclusões e extensões

6.1 Contribuições

O planejamento eficiente das atividades agrícolas requer monitoramentos e caracterizações dos elementos climáticos, alcançados através do processamento de séries históricas climatológicas. Neste sentido, a integração de dados de fontes diferentes possibilita identificações mais precisas dos padrões climáticos, melhorando a utilização dos recursos agrícolas e diminuindo perdas de safras. Os principais objetivos desta dissertação foram estudar os problemas inerentes à integração de dados climatológicos e propor uma abordagem que resolvesse tais questões.

Inicialmente estudou-se metodologias de integração de dados propostas na literatura. A seguir, foi analisada a situação corrente da disponibilização de dados climatológicos no Brasil. Esta análise foi beneficiada pelo contato com um caso real vivido na EMBRAPA, na qual também foram levantados requisitos de usuário para um sistema de dados climatológicos. A etapa seguinte consistiu em estudar técnicas de avaliação de qualidade, incluindo métodos específicos para séries históricas pluviométricas. Também foram pesquisados na literatura padrões de metadados que se relacionassem com dados climatológicos. Adicionalmente, foram discutidas questões referentes ao intercâmbio de dados na Web, campo de pesquisa importante quando se trabalha com compartilhamento de dados.

O resultado destes estudos foi a elaboração de uma arquitetura para um sistema de integração de dados climatológicos. A solução proposta é baseada na integração das séries históricas de diversas fontes em uma base de dados única. A especificação das consultas é facilitada por um banco de metadados, que descreve características relevantes dos dados, e por resultados de avaliações de qualidade realizadas sobre os dados. Estas avaliações também fornecem parâmetros para a credibilidade dos resultados obtidos, dando mais suporte às tomadas de decisão baseadas nas informações retornadas pelas consultas.

A arquitetura proposta foi implementada num caso real, que pôs em prática as metodologias de homogeneização de dados e de avaliação de qualidade. Foram desenvolvidos métodos de consistência de medidas pluviométricas, aplicados a séries históricas com décadas de dados.

Em resumo, as principais contribuições desta dissertação foram:

- estudo e levantamento dos problemas inerentes à integração de dados climatológicos heterogêneos;
- levantamento e discussão das questões envolvidas em avaliação de qualidade de dados;
- proposta de um conjunto de metadados para descrever dados climatológicos;
- proposta de uma arquitetura de integração de dados climatológicos;
- proposta de uma metodologia de avaliação de qualidade de dados; e
- implementação parcial da arquitetura e discussão das dificuldades encontradas na implementação.

A parte da arquitetura implementada foram o Módulo de Integração e o Módulo de Avaliação de Qualidade.

6.2 Extensões

As extensões previstas para esta dissertação são de dois tipos: extensões da arquitetura e extensões de implementação.

6.2.1 Extensões da arquitetura

Publicação de dados na Web - Uma das maneiras de disponibilizar bases de dados climatológicos para outras instituições é publicá-los na Web. Neste contexto, é necessário pesquisar maneiras de representar e formatar os dados para que a publicação seja facilmente assimilável por diversos sistemas. Uma possibilidade é utilizar XML para publicação de dados e RDF para a publicação de metadados, como discutido no capítulo 2.

Ontologias para dados climatológicos - Esta linha de pesquisa pode ser vista como uma extensão à anterior, uma vez que também visa aperfeiçoar o intercâmbio de dados. Como introduzido no capítulo 2, a modelagem de conceitos sob a forma de ontologias permite que sistemas interpretem a semântica de bases de dados. Neste sentido, o uso de ontologias pode trazer benefícios para o intercâmbio de dados climatológicos [35].

Mineração de dados em séries históricas - O uso de métodos de *data mining* possibilita descoberta de conhecimentos ocultos em grandes bases de dados. Uma área importante de pesquisa é o desenvolvimento de métodos de *data mining* específicos para dados climatológicos, visando melhor identificação de padrões climáticos [29, 14].

Integração visando espacialização de produtos - A ênfase da integração, nesta dissertação, foi relativa aos valores dos atributos descritivos. Uma extensão futura é a integração levando em consideração aspectos espaciais, inclusive quando há problemas na qualidade da informação espacial. Esta integração deverá ser proposta visando os produtos a serem gerados.

Avaliação da reputação da fonte de dados - Uma extensão à metodologia de avaliação de qualidade de dados climatológicos proposta é a inclusão de um novo indicador de qualidade, a *reputação* da fonte de dados. Esta informação seria útil para a escolha das fontes de dados a serem utilizadas no sistema integrado. Uma possibilidade é determinar a reputação da fonte de dados a partir dos resultados das avaliações de qualidade não-dependente da tarefa realizadas sobre todos os seus dados.

Consideração do dispositivo de coleta na avaliação de qualidade - Uma segunda extensão à avaliação de qualidade proposta, no caso de dados pluviométricos, é considerar que tipo de dispositivo (pluviômetro ou pluviógrafo) foi utilizado na coleta do dado. Como erros de leitura são menos prováveis quando se usa pluviógrafos, a avaliação de qualidade pode aplicar tratamentos diferenciados para dados que venham de um ou outro dispositivo, podendo, por exemplo, atribuir maior confiança a dados oriundos de pluviógrafos.

6.2.2 Extensões de implementação

Interação com SIG - A versão atual do Módulo de Processamento de Consultas não interage com nenhum SIG e, portanto, não permite visualização cartográfica. Uma extensão ao módulo, portanto, seria a implementação desta conexão.

Visualização das notas de qualidade em consultas - Uma segunda extensão ao Módulo de Processamento de Consultas seria a implementação da visualização das notas finais dos quatro indicadores de qualidade (completude, ausência de erro, atualidade e quantidade apropriada de dados) associadas às estações, em cada consulta. Pode-se estudar a criação de uma interface que permitisse ao usuário visualizar os resultados das avaliações, mesmo quando a consulta envolvesse muitas estações.

Cálculos de métrica de qualidade global - Finalmente, outra extensão é a implementação de uma maneira de o usuário especificar em tempo de execução a importância de cada indicador de qualidade. Isto permitiria o cálculo de uma só nota de qualidade para a consulta.

Uso de *data warehouses* - A arquitetura proposta é baseada em um SGBD relacional convencional. Uma possível extensão é a utilização de *data warehouses*, o que exige considerar aspectos como visões materializadas em que se armazenam valores agregados (como, por exemplo, médias, totais e extremos diários, semanais e mensais).

Apêndice A

Esquema do banco de metadados

A Figura A.1 exibe o esquema resultante da especificação do banco de metadados. A relação *Description* é a principal relação, e implementa a entidade *Descrição* do diagrama da Figura 4.3 (pág. 53). O atributo *id.description* desta relação corresponde ao elemento *Identifier* do conjunto de Metadados de Identificação (pág. 55). Ele consiste no identificador do conjunto de dados descrito, e é utilizado como chave de cada descrição armazenada no banco de metadados.

Os elementos monovalorados do conjunto de metadados são implementados como atributos da relação *Description*. A exceção é o elemento *Keywords*, do conjunto de Metadados de Identificação, que, apesar de ser multivalorado, é implementado como um único atributo da relação *Description*, no qual as palavras-chave que descrevem o conjunto de dados são armazenadas separadas por vírgula, para facilitar a indexação.

A relação *Creator* armazena, para cada descrição, os elementos multivalorados *Creator* e *Address* do conjunto de Metadados de Identificação.

A relação *Region* armazena o elemento multivalorado *Region* do conjunto de Metadados de Cobertura Espacial (pág. 56). Os outros elementos multivalorados deste conjunto, *Location*, *Latitude* e *Longitude*, são armazenados na relação *Location*.

A relação *Temporal_coverage* armazena o elemento multivalorado *Coverage.temporal* do conjunto de Metadados de Cobertura Temporal (pág. 56).

Description

id_description: VARCHAR2(20) NOT NULL

title: VARCHAR2(40) NOT NULL
description: VARCHAR2(80) NOT NULL
purpose: LONG VARCHAR NOT NULL
status: VARCHAR2(20) NOT NULL
update_frequency: VARCHAR2(20) NOT NULL
rights: VARCHAR2(20) NOT NULL
date_modified: DATE NULL
lineage: LONG VARCHAR NULL
format: VARCHAR2(20) NOT NULL
keywords: VARCHAR2(50) NOT NULL
relation_is_part_of: VARCHAR2(30) NULL
online_linkage: VARCHAR2(40) NULL
granularity: VARCHAR2(20) NOT NULL
latitude_unit: VARCHAR2(40) NOT NULL
longitude_unit: VARCHAR2(40) NOT NULL
latitude_resolution: VARCHAR2(15) NOT NULL
longitude_resolution: VARCHAR2(15) NOT NULL
altitude_unit: VARCHAR2(25) NOT NULL
precipitation_unit: VARCHAR2(25) NOT NULL
temperature_unit: VARCHAR2(25) NOT NULL
atmospheric_pressure_unit: VARCHAR2(25) NOT NULL
relative_humidity_unit: VARCHAR2(25) NOT NULL
radiation_unit: VARCHAR2(25) NOT NULL
responsible: VARCHAR2(40) NOT NULL
contact: VARCHAR2(100) NULL
language: VARCHAR2(5) NOT NULL
last_revision: DATE NULL
next_revision: DATE NULL

Creator

id_creator: NUMBER NOT NULL

id_description: VARCHAR2(20) NOT NULL
creator: VARCHAR2(50) NOT NULL
address: LONG VARCHAR NULL

Region

id_region: NUMBER NOT NULL

id_description: VARCHAR2(20) NOT NULL
region: VARCHAR2(30) NOT NULL

Location

id_location: NUMBER NOT NULL

id_description: VARCHAR2(20) NOT NULL
location: VARCHAR2(25) NOT NULL
latitude: NUMBER NOT NULL
longitude: NUMBER NULL

Temporal coverage

id_temporal_coverage: NUMBER NOT NULL

id_description: VARCHAR2(20) NOT NULL
initial_date: DATE NOT NULL
final_date: DATE NOT NULL

Figura A.1: Esquema do banco de metadados.

Apêndice B

Esquema XML para os dados homogeneizados

A seguir é apresentado o esquema em XML Schema utilizado para validar os documentos XML gerados pelos tradutores de dados. O esquema define que os documentos XML devem ser compostos por duas partes, sendo a primeira dedicada aos metadados e a segunda aos dados climatológicos.

A parte relacionada aos metadados deve consistir de um ou mais elementos `description`, que, por sua vez, devem ser formados por uma sequência dos elementos `identificationMetadata`, `spatialCoverageMetadata`, `temporalCoverageMetadata`, `unitsMetadata` e `managementMetadata`.

A parte do documento relacionada aos dados climatológicos deve ser composta por um elemento `institution`, seguido por um ou mais elementos `station`, seguidos por um ou mais elementos `hourlyMeasure` e `dailyMeasure`. Estes últimos devem referenciar, através de um de seus atributos, um dos elementos `description` existentes no documento. Isto garante que todas as medidas constantes no documento se relacionem com uma descrição nos metadados.

O esquema encontra-se disponível em <http://www.ic.unicamp.br/proj-adb/climatological/climatologicalSchema#>.

```

<?xml version="1.0" encoding="UTF-8"?>
<xs:schema targetNamespace="http://www.ic.unicamp.br/proj-adb/climatological/climatologicalSchema#"
xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:tns="http://www.ic.unicamp.br/proj-
adb/climatological/climatologicalSchema#" elementFormDefault="qualified" attributeFormDefault="unqualified">
  <!-- -->
  <!--Metadata types.-->
  <!-- -->
  <xs:complexType name="creatorType">
    <xs:sequence>
      <xs:element name="name" type="xs:token"/>
      <xs:element name="address" type="xs:string" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="countyType">
    <xs:sequence>
      <xs:element name="name" type="xs:token"/>
      <xs:element name="latitude" type="tns:latitudeType"/>
      <xs:element name="longitude" type="tns:longitudeType"/>
    </xs:sequence>
  </xs:complexType>
  <xs:complexType name="dateRangeType">
    <xs:sequence>
      <xs:element name="initialDate" type="xs:date"/>
      <xs:element name="finalDate" type="xs:date"/>
    </xs:sequence>
  </xs:complexType>
  <!-- -->
  <xs:complexType name="identificationMetadataType">
    <xs:sequence>
      <xs:element name="title">
        <xs:simpleType>
          <xs:restriction base="xs:token">
            <xs:maxLength value="60"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element name="descriptionPhrase" type="xs:string"/>
      <xs:element name="purpose" type="xs:string"/>
      <xs:element name="status">
        <xs:simpleType>
          <xs:restriction base="xs:token">
            <xs:maxLength value="20"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element name="updateFrequency">
        <xs:simpleType>
          <xs:restriction base="xs:token">
            <xs:maxLength value="20"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element name="rights" type="xs:string"/>
      <xs:element name="creator" type="tns:creatorType" maxOccurs="unbounded"/>
      <xs:element name="dateModified" type="xs:date" minOccurs="0"/>
      <xs:element name="lineage" type="xs:string" minOccurs="0"/>
      <xs:element name="format" type="xs:string" minOccurs="0"/>
      <xs:element name="keywords" type="xs:token"/>
      <xs:element name="relationsPartOf" minOccurs="0">
        <xs:simpleType>
          <xs:restriction base="xs:token">
            <xs:maxLength value="30"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
    </xs:sequence>
  </xs:complexType>

```

```

        </xs:restriction>
      </xs:simpleType>
    </xs:element>
    <xs:element name="onlineLinkage" type="xs:Name" minOccurs="0"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="spatialCoverageMetadataType">
  <xs:sequence>
    <xs:element name="region" type="xs:token" minOccurs="0" maxOccurs="unbounded"/>
    <xs:element name="county" type="tns:countyType" maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="temporalCoverageMetadataType">
  <xs:sequence>
    <xs:element name="granularity" type="xs:token"/>
    <xs:element name="dateRange" type="tns:dateRangeType" maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="unitsMetadataType">
  <xs:sequence>
    <xs:element name="latitudeUnit" type="xs:token"/>
    <xs:element name="longitudeUnit" type="xs:token"/>
    <xs:element name="latitudeResolution" type="xs:token"/>
    <xs:element name="longitudeResolution" type="xs:token"/>
    <xs:element name="altitudeUnit" type="xs:token"/>
    <xs:element name="precipitationUnit" type="xs:token" minOccurs="0"/>
    <xs:element name="temperatureUnit" type="xs:token" minOccurs="0"/>
    <xs:element name="atmosphericPressureUnit" type="xs:token" minOccurs="0"/>
    <xs:element name="relativeHumidityUnit" type="xs:token" minOccurs="0"/>
    <xs:element name="radiationUnit" type="xs:token" minOccurs="0"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="managementMetadataType">
  <xs:sequence>
    <xs:element name="responsible" type="xs:token"/>
    <xs:element name="contact" type="xs:string" minOccurs="0"/>
    <xs:element name="language" type="xs:language"/>
    <xs:element name="lastRevision" type="xs:date" minOccurs="0"/>
    <xs:element name="nextRevision" type="xs:date" minOccurs="0"/>
  </xs:sequence>
</xs:complexType>
<!-- -->
<xs:complexType name="descriptionType">
  <xs:sequence>
    <xs:element name="identificationMetadata" type="tns:identificationMetadataType"/>
    <xs:element name="spatialCoverageMetadata" type="tns:spatialCoverageMetadataType"/>
    <xs:element name="temporalCoverageMetadata" type="tns:temporalCoverageMetadataType"/>
    <xs:element name="unitsMetadata" type="tns:unitsMetadataType"/>
    <xs:element name="managementMetadata" type="tns:managementMetadataType"/>
  </xs:sequence>
  <xs:attribute name="idDescription" type="xs:integer" use="required"/>
</xs:complexType>
<xs:complexType name="climatologicalMetadataType">
  <xs:sequence>
    <xs:element name="description" type="tns:descriptionType" maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<!-- -->
<!--Data types.-->
<!-- -->
<xs:simpleType name="phoneType">
  <xs:restriction base="xs:string">

```



```

    <xs:pattern value="\{[0-9][0-9][0-9]{4}-[0-9]{4}\}/>
    <xs:pattern value="\{[0-9][0-9][0-9]{3}-[0-9]{4}\}/>
  </xs:restriction>
</xs:simpleType>
<xs:simpleType name="emailType">
  <xs:restriction base="xs:string">
    <xs:pattern value="(lc)+@{(lc)+}/>
  </xs:restriction>
</xs:simpleType>
<xs:simpleType name="latitudeType">
  <xs:restriction base="xs:string"/>
</xs:simpleType>
<xs:simpleType name="longitudeType">
  <xs:restriction base="xs:string"/>
</xs:simpleType>
<xs:simpleType name="altitudeType">
  <xs:restriction base="xs:decimal"/>
</xs:simpleType>
<xs:simpleType name="stateType">
  <xs:restriction base="xs:string">
    <xs:pattern value="[A-Z]{2}/>
  </xs:restriction>
</xs:simpleType>
<xs:simpleType name="temperatureType">
  <xs:restriction base="xs:decimal"/>
</xs:simpleType>
<xs:simpleType name="relativeHumidityType">
  <xs:restriction base="xs:decimal">
    <xs:minInclusive value="0"/>
  </xs:restriction>
</xs:simpleType>
<xs:simpleType name="precipitationType">
  <xs:restriction base="xs:decimal"/>
</xs:simpleType>
<xs:simpleType name="solarRadiationType">
  <xs:restriction base="xs:decimal">
    <xs:minInclusive value="0"/>
  </xs:restriction>
</xs:simpleType>
<xs:simpleType name="atmosphericPressureType">
  <xs:restriction base="xs:decimal"/>
</xs:simpleType>
<xs:simpleType name="idStationType">
  <xs:restriction base="xs:string"/>
</xs:simpleType>
<!-- -->
<xs:complexType name="institutionType">
  <xs:sequence>
    <xs:element name="acronym">
      <xs:simpleType>
        <xs:restriction base="xs:string">
          <xs:pattern value="[A-Z]+"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:element>
    <xs:element name="name" type="xs:token"/>
    <xs:element name="address" type="xs:string" minOccurs="0"/>
    <xs:element name="phone" type="tns:phoneType" minOccurs="0"/>
    <xs:element name="email" type="tns:emailType" minOccurs="0"/>
    <xs:element name="URL" type="xs:Name" minOccurs="0"/>
    <xs:element name="contact" type="xs:string" minOccurs="0"/>
  </xs:sequence>

```

```

</xs:complexType>
<xs:complexType name="stationType">
  <xs:sequence>
    <xs:element name="name" type="xs:token" minOccurs="0"/>
    <xs:element name="latitude" type="tns:latitudeType"/>
    <xs:element name="longitude" type="tns:longitudeType"/>
    <xs:element name="altitude" type="tns:altitudeType"/>
    <xs:element name="county" type="xs:token"/>
    <xs:element name="state" type="tns:stateType"/>
    <xs:element name="address" type="xs:string" minOccurs="0"/>
    <xs:element name="phone" type="tns:phoneType" minOccurs="0"/>
    <xs:element name="contact" type="xs:string" minOccurs="0"/>
  </xs:sequence>
  <xs:attribute name="idStation" type="tns:idStationType" use="required"/>
</xs:complexType>
<xs:complexType name="hourlyMeasureType">
  <xs:all>
    <xs:element name="temperature" type="tns:temperatureType" minOccurs="0"/>
    <xs:element name="relativeHumidity" type="tns:relativeHumidityType" minOccurs="0"/>
    <xs:element name="accumulatedPrecipitation" type="tns:precipitationType" minOccurs="0"/>
    <xs:element name="solarRadiation" type="tns:solarRadiationType" minOccurs="0"/>
    <xs:element name="atmosphericPressure" type="tns:atmosphericPressureType" minOccurs="0"/>
  </xs:all>
  <xs:attribute name="idStation" type="tns:idStationType" use="required"/>
  <xs:attribute name="date" type="xs:date" use="required"/>
  <xs:attribute name="time" type="xs:time" use="required"/>
  <xs:attribute name="idDescription" type="xs:integer" use="optional"/>
</xs:complexType>
<xs:complexType name="dailyMeasureType">
  <xs:all>
    <xs:element name="maximumTemperature" type="tns:temperatureType" minOccurs="0"/>
    <xs:element name="minimumTemperature" type="tns:temperatureType" minOccurs="0"/>
    <xs:element name="minimumRelativeHumidity" type="tns:relativeHumidityType" minOccurs="0"/>
    <xs:element name="mediumRelativeHumidity" type="tns:relativeHumidityType" minOccurs="0"/>
    <xs:element name="maximumRelativeHumidity" type="tns:relativeHumidityType" minOccurs="0"/>
    <xs:element name="accumulatedPrecipitation" type="tns:precipitationType" minOccurs="0"/>
    <xs:element name="accumulatedSolarRadiation" type="tns:solarRadiationType" minOccurs="0"/>
  </xs:all>
  <xs:attribute name="idStation" type="tns:idStationType" use="required"/>
  <xs:attribute name="date" type="xs:date" use="required"/>
  <xs:attribute name="idDescription" type="xs:integer" use="optional"/>
</xs:complexType>
<!-- -->
<xs:complexType name="climatologicalDataType">
  <xs:sequence>
    <xs:element name="institution" type="tns:institutionType"/>
    <xs:element name="station" type="tns:stationType" maxOccurs="unbounded"/>
    <xs:choice maxOccurs="unbounded">
      <xs:element name="dailyMeasure" type="tns:dailyMeasureType"/>
      <xs:element name="hourlyMeasure" type="tns:hourlyMeasureType"/>
    </xs:choice>
  </xs:sequence>
</xs:complexType>
<!-- -->
<!--Root element type.-->
<!-- -->
<xs:complexType name="climatologicalDataAndMetadataType">
  <xs:sequence>
    <xs:element name="climatologicalMetadata" type="tns:climatologicalMetadataType"/>
    <xs:element name="climatologicalData" type="tns:climatologicalDataType"/>
  </xs:sequence>
</xs:complexType>

```

```

<!-- -->
<!-- Root element.-->
<!-- -->
<xs:element name="climatologicalDataAndMetadata" type="tns:climatologicalDataAndMetadataType">
  <!-- -->
  <!-- Description key.-->
  <!-- -->
  <xs:key name="descriptionKey">
    <xs:selector xpath="tns:climatologicalMetadata/tns:description"/>
    <xs:field xpath="@idDescription"/>
  </xs:key>
  <!-- -->
  <!-- Station key.-->
  <!-- -->
  <xs:key name="stationKey">
    <xs:selector xpath="tns:climatologicalData/tns:station"/>
    <xs:field xpath="@idStation"/>
  </xs:key>
  <!-- -->
  <!-- Each daily measure must refer one station.-->
  <!-- -->
  <xs:keyref name="dailyMeasureRefersStation" refer="tns:stationKey">
    <xs:selector xpath="tns:climatologicalData/tns:dailyMeasure"/>
    <xs:field xpath="@idStation"/>
  </xs:keyref>
  <!-- -->
  <!-- Each daily measure must refer one description.-->
  <!-- -->
  <xs:keyref name="dailyMeasureRefersDescription" refer="tns:descriptionKey">
    <xs:selector xpath="tns:climatologicalData/tns:dailyMeasure"/>
    <xs:field xpath="@idDescription"/>
  </xs:keyref>
  <!-- -->
  <!-- Each hourly measure must refer one station.-->
  <!-- -->
  <xs:keyref name="hourlyMeasureRefersStation" refer="tns:stationKey">
    <xs:selector xpath="tns:climatologicalData/tns:hourlyMeasure"/>
    <xs:field xpath="@idStation"/>
  </xs:keyref>
  <!-- -->
  <!-- Each hourly measure must refer one description.-->
  <!-- -->
  <xs:keyref name="hourlyMeasureRefersDescription" refer="tns:descriptionKey">
    <xs:selector xpath="tns:climatologicalData/tns:hourlyMeasure"/>
    <xs:field xpath="@idDescription"/>
  </xs:keyref>
</xs:element>
<!-- -->
</xs:schema>

```

A seguir é apresentado um exemplo de documento XML válido. Este exemplo contém dados de chuva, temperatura e umidade relativa do ar de duas estações climatológicas, dos municípios de Uberaba e Uberlândia, no período de 01/09/2002 a 03/09/2002.

```

<?xml version="1.0" encoding="UTF-8"?>
<climatologicalDataAndMetadata xmlns="http://www.ic.unicamp.br/proj-
adb/climatological/climatologicalSchema#" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.ic.unicamp.br/proj-adb/climatological/climatologicalSchema#">
  <climatologicalMetadata>
    <description idDescription="001">
      <identificationMetadata>
        <title>Dados de Uberlândia e Uberaba</title>
        <descriptionPhrase>Dados pluviométricos diários e horários, dados de temperatura diários e
dados de umidade diários.</descriptionPhrase>
        <purpose>Geração de estatísticas e previsões climáticas de chuva, temperatura e umidade.
Devido à sua extensão limitada, estes dados devem usados preferencialmente em conjunto com outros
dados da mesma região geográfica.</purpose>
        <status>em modificação</status>
        <updateFrequency>horária</updateFrequency>
        <rights>O uso destes dados é livre.</rights>
        <creator>
          <name>Companhia Climatológica Brasileira</name>
          <address>Av. das Nações Unidas, 1450, 3º andar, São Paulo, SP. CEP 11010-000</address>
        </creator>
        <dateModified>2002-09-16</dateModified>
        <lineage>Os dados foram coletados por instrumentos mecânicos, lidos e transcritos por
operadores humanos. Os dados não passaram por nenhum tipo de consistência.</lineage>
        <keywords>dados climatológicos,dados
pluviométricos,hidrologia,temperatura,umidade</keywords>
      </identificationMetadata>
      <spatialCoverageMetadata>
        <region>Sudeste de Minas Gerais.</region>
        <county>
          <name>Uberlandia</name>
          <latitude>-25.09</latitude>
          <longitude>-100.00</longitude>
        </county>
        <county>
          <name>Uberaba</name>
          <latitude>-28.3</latitude>
          <longitude>-110.4</longitude>
        </county>
      </spatialCoverageMetadata>
      <temporalCoverageMetadata>
        <granularity>horária</granularity>
        <dateRange>
          <initialDate>2002-09-01</initialDate>
          <finalDate>2002-09-03</finalDate>
        </dateRange>
      </temporalCoverageMetadata>
      <unitsMetadata>
        <latitudeUnit>Graus decimais</latitudeUnit>
        <longitudeUnit>Graus decimais</longitudeUnit>
        <latitudeResolution>0.01</latitudeResolution>
        <longitudeResolution>0.01</longitudeResolution>
        <altitudeUnit>Metros</altitudeUnit>
        <precipitationUnit>Milímetros</precipitationUnit>
        <temperatureUnit>Graus Celsius</temperatureUnit>
        <relativeHumidityUnit>Porcentagem</relativeHumidityUnit>
      </unitsMetadata>
      <managementMetadata>
        <responsible>Sílvio</responsible>
        <contact>(22)5555-4687 silvio@ccb.gov.br</contact>
        <language>pt</language>
        <lastRevision>2002-09-30</lastRevision>
      </managementMetadata>
    </description idDescription="001">
  </climatologicalMetadata>
</climatologicalDataAndMetadata>

```

```

</description>
</climatologicalMetadata>
<climatologicalData>
  <institution>
    <acronym>CCB</acronym>
    <name>Companhia Climatológica Brasileira</name>
    <address>Rua da Saudade, 100</address>
    <phone>(22)5555-4546</phone>
    <email>ccb@ccb.gov.br</email>
    <contact>Sílvio: (22)5555-4687 silvio@ccb.gov.br</contact>
  </institution>
  <station idStation="estacao001">
    <name>Uberlândia</name>
    <latitude>-25.09</latitude>
    <longitude>-100.00</longitude>
    <altitude>134</altitude>
    <county>Uberlândia</county>
    <state>MG</state>
    <address>Estrada da Fé, km 34</address>
  </station>
  <station idStation="estacao002">
    <name>Uberaba</name>
    <latitude>-28.3</latitude>
    <longitude>-110.4</longitude>
    <altitude>25.4</altitude>
    <county>Uberlândia</county>
    <state>SP</state>
  </station>
  <hourlyMeasure idStation="estacao001" date="2002-09-01" time="08:00" idDescription="001">
    <accumulatedPrecipitation>0</accumulatedPrecipitation>
  </hourlyMeasure>
  <hourlyMeasure idStation="estacao001" date="2002-09-01" time="09:00" idDescription="001">
    <accumulatedPrecipitation>12.3</accumulatedPrecipitation>
  </hourlyMeasure>
  <hourlyMeasure idStation="estacao001" date="2002-09-01" time="10:00" idDescription="001">
    <accumulatedPrecipitation>15</accumulatedPrecipitation>
  </hourlyMeasure>
  <hourlyMeasure idStation="estacao002" date="2002-09-01" time="08:00" idDescription="001">
    <accumulatedPrecipitation>0</accumulatedPrecipitation>
  </hourlyMeasure>
  <hourlyMeasure idStation="estacao002" date="2002-09-01" time="09:00" idDescription="001">
    <accumulatedPrecipitation>0</accumulatedPrecipitation>
  </hourlyMeasure>
  <hourlyMeasure idStation="estacao002" date="2002-09-01" time="10:00" idDescription="001">
    <accumulatedPrecipitation>0</accumulatedPrecipitation>
  </hourlyMeasure>
  <dailyMeasure idStation="estacao001" date="2002-09-01" idDescription="001">
    <maximumTemperature>24</maximumTemperature>
    <minimumTemperature>18</minimumTemperature>
    <mediumRelativeHumidity>67.8</mediumRelativeHumidity>
    <accumulatedPrecipitation>26</accumulatedPrecipitation>
  </dailyMeasure>
  <dailyMeasure idStation="estacao001" date="2002-09-02" idDescription="001">
    <maximumTemperature>28</maximumTemperature>
    <minimumTemperature>20</minimumTemperature>
    <mediumRelativeHumidity>70</mediumRelativeHumidity>
    <accumulatedPrecipitation>12</accumulatedPrecipitation>
  </dailyMeasure>
  <dailyMeasure idStation="estacao001" date="2002-09-03" idDescription="001">
    <maximumTemperature>25</maximumTemperature>
    <minimumTemperature>17</minimumTemperature>
    <mediumRelativeHumidity>66.8</mediumRelativeHumidity>
  </dailyMeasure>

```

```
<accumulatedPrecipitation>0</accumulatedPrecipitation>
</dailyMeasure>
<dailyMeasure idStation="estacao002" date="2002-09-01" idDescription="001">
  <maximumTemperature>28</maximumTemperature>
  <minimumTemperature>22</minimumTemperature>
  <mediumRelativeHumidity>56.8</mediumRelativeHumidity>
  <accumulatedPrecipitation>0</accumulatedPrecipitation>
</dailyMeasure>
<dailyMeasure idStation="estacao002" date="2002-09-02" idDescription="001">
  <maximumTemperature>32</maximumTemperature>
  <minimumTemperature>24</minimumTemperature>
  <mediumRelativeHumidity>87.7</mediumRelativeHumidity>
  <accumulatedPrecipitation>45</accumulatedPrecipitation>
</dailyMeasure>
<dailyMeasure idStation="estacao002" date="2002-09-03" idDescription="001">
  <maximumTemperature>34</maximumTemperature>
  <minimumTemperature>22</minimumTemperature>
  <mediumRelativeHumidity>82.3</mediumRelativeHumidity>
  <accumulatedPrecipitation>15</accumulatedPrecipitation>
</dailyMeasure>
</climatologicalData>
</climatologicalDataAndMetadata>
```

Referências Bibliográficas

- [1] *Dublin Core Metadata Element Set*. <http://dublincore.org/documents/dces> (consultado em novembro de 2003).
- [2] *Dublin Core Metadata Initiative*. <http://dublincore.org> (consultado em novembro de 2003).
- [3] PIB do agronegócio brasileiro chega a 424,32 bilhões de reais. *FAEP - Federação da Agricultura do Estado do Paraná*. Boletim Informativo 760, semana de 17 a 23 de março de 2003.
- [4] Rain gauge consistency check using the double-mass curve technique - example. Technical report, School of Aeronautical, Civil and Mechanical Engineering - University of Salford, Salford, Reino Unido.
- [5] *World Meteorological Organization*. <http://www.wmo.ch> (consultado em maio de 2003).
- [6] H. Alvestrand. *RFC3066 - Tags for the Identification of Languages*. <http://www.ietf.org/rfc/rfc3066.txt> (consultado em junho de 2003).
- [7] G. Aslan and D. Leod. Semantic heterogeneity resolution in federated databases by metadata implantation and stepwise evolution. *The VLDB Journal*, 8(2):120–132, 1999.
- [8] E. N. Australia. *EdNA Metadata Standard v1.1*. <http://www.edna.edu.au/metadata> (consultado em março de 2003).
- [9] B. Bachimont, A. Isaac, and R. Troncy. Semantic commitment for designing ontologies: a proposal. *EKAW 2002, Lecture Notes in Artificial Intelligence (LNAI)*, (2473):114–121, 2002.
- [10] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, pages 34–43, maio 2001.

- [11] T. Bray. *What is RDF?*. xml.com, <http://www.xml.com/pub/a/2001/01/24/rdf.html> (consultado em janeiro de 2003).
- [12] D. Brickley and R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*. <http://www.w3.org/TR/rdf-schema> (consultado em janeiro de 2003).
- [13] G. Camara, M. A. Casanova, A. S. Hemerly, G. C. Magalhães, and C. B. Medeiros. *Anatomia de Sistemas de Informação Geográfica*. Instituto de Computação, UNICAMP, 1996.
- [14] P. Carbone. Data mining or knowledge discovery in databases: An overview. In *Data Management Handbook*, New York, USA, 1997.
- [15] N. Cereja. Visões em sistemas de informação geográficos - modelo e mecanismos. Master's thesis, Instituto de Computação, UNICAMP, 1996.
- [16] N. R. Chrisman. Living with error in geographic data: Truth and responsibility. In *Proceedings GIS*, volume 1, pages 12–17, Vancouver BC, Canadá, 1995.
- [17] D. Connolly and T. Berners-Lee. *Naming and Addressing: URIs, URLs, ...* <http://www.w3.org/Addressing/> (consultado em junho de 2003).
- [18] O. G. Consortium. *Geospatial Interoperability: the Open GIS Consortium Perspective*. <http://www.opengis.org> (consultado em janeiro de 2003).
- [19] W. W. W. Consortium. *Extensible Markup Language (XML)*. <http://www.w3.org/XML> (consultado em janeiro de 2003).
- [20] W. W. W. Consortium. *Feature Synopsis for OWL Lite and OWL*. <http://www.w3.org/TR/owl-features> (consultado em janeiro de 2003).
- [21] M. Corporation. *Microsoft Visual Basic for Applications Home Page*. <http://msdn.microsoft.com/vba> (consultado em julho de 2003).
- [22] O. Corporation. *Current Oracle Documentation*. <http://otn.oracle.com/documentation> (consultado em julho de 2003).
- [23] R. Costanza, S. O. Funtowicz, and J. R. Ravetz. Assessing and communicating data quality in policy-relevant research. *Environmental Management*, 16(1):121–131, 1992.
- [24] A. C. de Alencar. Qualidade de dados em aplicações geográficas. Master's thesis, Instituto de Computação, UNICAMP, 2000.

- [25] J. G. de S. Lima, C. B. Medeiros, and E. D. Assad. Integration of heterogeneous pluviometric data for crop forecasts. In *V Simpósio Brasileiro de Geoinformática - GEOINFO*, Campos do Jordão, 2003.
- [26] D. Defense Advanced Research Projects Agency. *The DARPA Agent Markup Language*. <http://www.daml.org> (consultado em janeiro de 2003).
- [27] C. B. M. e F. Pires. Databases for GIS. *ACM SIGMOD Record*, 23(1):107–115, 1994.
- [28] D. Fallside. *XML-Schema Part 0: Primer*. <http://www.w3.org/TR/2001/REC-xmlschema-0-20010502> (consultado em janeiro de 2003).
- [29] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *Ai Magazine*, 17:37–54, 1996.
- [30] U. G. Federal Geographic Data Committee. *Content Standards for Digital Geospatial Metadata*. <http://www.fgdc.gov/metadata/contstan.html> (consultado em março de 2003).
- [31] C. Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, 1998.
- [32] D. Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, 2001.
- [33] D. Fensel. Ontology-Based Knowledge Management. *IEEE Computer*, 35(11):56–59, 2002.
- [34] D. Fensel and M. A. Musen. The semantic web: a brain for humankind. *IEEE Intelligent Systems*, 16(2):24–25, 2001.
- [35] R. Fileto. Issues on interoperability and integration of heterogeneous geographical data. In *III Workshop Brasileiro de Geoinformática - GEOINFO*, pages 133–140, Rio de Janeiro, 2001.
- [36] F. T. Fonseca. Role-based geographic information integration. In *III Workshop Brasileiro de Geoinformática - GEOINFO*, pages 31–38, Rio de Janeiro, 2001.
- [37] Food and A. O. of the United Nations. *The Agricultural Metadata Standards Initiative*. <http://www.fao.org/agris/MagazineArchive/MetaData/TaskForceonDCMI.htm> (consultado em março de 2003).
- [38] I. Horrocks, D. Fensel, J. Broekstra, S. Decker, M. Erdmann, C. Goble, F. van Harmelen, M. Klein, S. Staab, R. Studer, and E. Motta. OIL: The Ontology Inference

- Layer. Technical Report IR-479, Vrije Universiteit Amsterdam, Faculty of Sciences, setembro 2000.
- [39] G. I. Institute. *Thesaurus of Geographic Names*. http://www.getty.edu/research/conducting_research/vocabularies/tgn/about.html (consultado em novembro de 2003).
- [40] M. Klein. XML, RDF, and relatives. *IEEE Intelligent Systems*, 16(2):26–28, 2001.
- [41] M. Klein, D. Fensel, F. van Harmelen, and I. Horrocks. The relation between ontologies and schema-languages: Translating OIL-specifications in XML-Schema. In *ECAI Workshop on Applications of Ontologies and Problem-Solving Methods*, Amsterdam, Netherlands, 2000. IOS Press.
- [42] D. S. Linthicum. Remember the metadata. *eAI Journal*, (9):8–10, setembro 2002.
- [43] C. López, E. González, and J. Goyret. Análisis por componentes principales de datos pluviométricos. a) aplicación a la detección de datos anómalos. *Estadística (Journal of the Inter-American Statistical Institute)*, (46):25–54, 1994.
- [44] C. López, J. F. González, and R. Curbelo. Análisis por componentes principales de datos pluviométricos. b) aplicación a la eliminación de ausencias. *Estadística (Journal of the Inter-American Statistical Institute)*, (46):55–83, 1994.
- [45] D. A. Mead. Assessing data quality in geographic information systems. In C. J. Johannsen and J. L. Sanders, editors, *Remote Sensing for Resource Management*, chapter 5, pages 51–62. Soil Conservation Society of America, 1982.
- [46] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini. Managing data quality in cooperative information systems. *10° Italian Symposium on Advanced Database Systems - SEBD 2002*, 2002.
- [47] C. B. Medeiros. Bancos de dados espaço-temporais: Fundamentos e aplicações. In *VI Escola Regional de Informática - Anais*, pages 241–255, ICMC-USP, São Carlos, 2001.
- [48] C. B. Medeiros and A. C. de Alencar. Qualidade dos dados e interoperabilidade em SIG. In *I Workshop Brasileiro de Geoinformática - GEOINFO*, pages 45–49, Campinas, 1999.
- [49] S. Microsystems. *The Source for Java Technology*. <http://java.sun.com/> (consultado em novembro de 2003).

- [50] E. Miller, R. Swick, and D. Brickley. *Resource Description Framework (RDF)*. <http://www.w3.org/RDF> (consultado em janeiro de 2003).
- [51] M. Missikoff, R. Navigli, and P. Velardi. Integrated approach to web ontology learning and engineering. *IEEE Computer*, 35(11):60–63, 2002.
- [52] National Center for Geographic Information e Analysis. *Interop. International Conference and Workshop on Interoperating Geographic Information Systems*, <http://www.ncgia.ucsb.edu/conf/interop97> (consultado em setembro de 2001).
- [53] I. Onyancha, F. Ward, F. Fisseha, K. Caprazli, S. Anibaldi, K. Johannes, and S. Katz. Metadata framework for resource discovery of agricultural information. In *Open Archives Initiative Workshop, 5th European Conference on Research and Advanced Technology for Digital Libraries*, Darmstadt, Alemanha, 2001.
- [54] W. M. Organization. *WMO Core Metadata Standard*. <http://www.wmo.ch/web/www/metadata/WMO-core-metadata-toc.html> (consultado em maio de 2003).
- [55] K. Orr. Data quality and systems theory. *Communications of the ACM*, 41(2):66–71, 1998.
- [56] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.
- [57] X. Qin. A case-based reasoning system for bearing design. Master's thesis, Drexel University, Philadelphia, EUA, 1999.
- [58] T. C. Redman. *Data Quality for the Information Age*. Artech House, 1996.
- [59] H. A. Rocha. Metadados para workflows científicos no apoio ao planejamento ambiental. Master's thesis, Instituto de Computação, UNICAMP, 2003.
- [60] S. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach*. Prentice-Hall, 1995.
- [61] A. Sheth and J. Larson. Federated database systems for managing distributed, heterogeneous and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, 1990.
- [62] H. R. Soares and C. B. Medeiros. Integrando sistemas legados a bancos de dados heterogêneos. *XIV Simpósio Brasileiro de Bancos de Dados*, 1:411–425, 1999.

- [63] S. Staab and A. Maedche. Ontology engineering beyond the modeling of concepts and relations. In *ECAI Workshop on Application of Ontologies and Problem-Solving Methods.*, Amsterdam, Holanda, 2000. IOS Press.
- [64] C. E. M. Tucci, editor. *Hidrologia: ciência e aplicação*, volume 4 of *Coleção ABRH de Recursos Hídricos*. Editora da Universidade, Porto Alegre, 1997.
- [65] M. Uschold and M. Gruninger. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.
- [66] F. Wang and S. Jusoh. Integrating multiple web-based geographic information systems. *IEEE Multimedia*, 6(1):49–61, 1999.
- [67] M. Wolf and C. Wicksteed. *World Wide Web Consortium - Date and Time Formats*. <http://www.w3.org/TR/NOTE-datetime> (consultado em junho de 2003).
- [68] M. Worboys and S. Deen. Semantic heterogeneity in distributed geographic databases. *ACM SIGMOD Record*, 20(4):30–34, 1991.