Extração de Dados em Sistemas de Informações Ambientais: Arquitetura e Esquema de Metadados

Hélio Alvaro de Mello Perez[1]
Ana Maria de Carvalho Moura[1]
Astério Kiyoshi Tanaka[2]

[1]Instituto Militar de Engenharia
Depto de Engenharia de Sistemas
e-mail : helio.perez@usa.net
anamoura@ime.eb.br

[2]Universidade do Rio de Janeiro
Depto de Informática Aplicada
e-mail : tanaka@uniriotec.br

Abstract. Environmental information is very important for the creation of environmental protection strategies, representing one of the greatest worries of most public organizations and societies in the world. Characterized by its heterogeneous and distributed nature, this kind of information needs to be integrated and changed into useful information, originating the so-called Environmental Information Systems (EIS). One of the greatest difficulties in constructing an EIS concerns, besides the huge amount of available data, the lack of a storage standard format, turning out its use and integration to be very difficult. Besides, the most specific and well-known metadata standards used for cataloguing environmental data still present a low power level for representing the structural part of data repositories. This paper presents a three-layer architecture to support the access and extraction processes of environmental data captured from heterogeneous and distributed repositories, and a metamodel based on metadata to enable the structure mapping of these data into an intermediate level, allowing users to build their programs to access the information directly, without changing the original storage location nor the data structure or data formats.

1. Introdução

Informações ambientais são imprescindíveis para a criação de estratégias de proteção ambiental. O problema é que, apesar da grande quantidade de dados disponíveis, não existe um formato padrão (planilhas, arquivos, banco de dados etc.) de armazenamento, dificultando sua integração e utilização pelas pessoas envolvidas no processo de tomada de decisão. Além disso, dados ambientais possuem características que dificultam o gerenciamento dos dados, tais como: grande volume de dados a ser processado; heterogeneidade de tipos de dados e de ambientes (software e hardware); estrutura complexa, com características espaço-temporal, incertos e altamente distribuídos (cada informação ou conjunto de dados pode estar localizada em locais diferentes, dificultando ainda mais sua recuperação e integração). Somam-se a isto ainda os principais problemas enfrentados pelos usuários na busca de informações ambientais [8]: os dados não existem ou são insuficientes; os dados não são referenciados pelos seus fornecedores ou são referenciados sob um determinado critério de classificação específico de um domínio; os dados são difíceis de serem acessados (precisam ser integrados e transformados); os conjuntos de dados acessados são difíceis de serem utilizados pois são inconsistentes ou incompatíveis; a qualidade dos dados recuperados é difícil de ser medida já que é difícil se comparar dados produzidos por diferentes modelos científicos sem sua documentação.

Apesar do grande número de soluções sendo utilizadas e em desenvolvimento, ainda não existe uma estrutura conceitual de integração destas soluções. As abordagens existentes são superficiais em relação aos tipos de repositórios de dados utilizados por um Sistema de Informação Ambiental (SIA), sendo necessárias transformações entre formatos, aumentando a complexidade das soluções.

Dentre as abordagens existentes, a mais viável são os padrões de metadados. Existem padrões específicos para a catalogação de dados ambientais ou geográficos, a exemplo do FGDC [2] e UDK [9]. Porém, durante o desenvolvimento desse trabalho foi verificado que, apesar desses padrões serem muito utilizados para a definição de informações ambientais, ainda possuem um baixo poder de representação da parte estrutural dos repositórios de dados. Como exemplo pode-se citar a quinta seção do padrão de metadados FGDC (seção 3.2 de [6]), que só

[1] Hélio Perez é consultor na área de administração de dados e de banco de dados. Atualmente presta consultoria no portal financeiro InvestShop.com http://www.investshop.com.br

[2] Professor do Departamento de Informática Aplicada da UNIRIO (Universidade do Rio de Janeiro) e Pesquisador Colaborador do Departamento de Engenharia de Sistemas do IME. Parcialmente patrocinado pelos projetos CNPq processo n.350652/94-5 e PROTEM-CC/INRIA processo n. 68.0139/98.2

contempla tabelas relacionais. Esta seção é responsável pela descrição do conteúdo da informação do conjunto de dados a nível de entidades, atributos e domínios, o que é ainda pouco abrangente em relação aos diversos tipos de repositórios existentes nos SIAs.

Além disso, a complexidade de um SIA é tão grande que a tendência é não permitir a criação de novos mecanismos para a conversão e tratamento de cada repositório de dados para um novo formato, o que geraria alguns problemas, tais como: gasto de espaço físico de armazenamento desnecessário e até mesmo inviável, devido ao volume de informações; direito de acesso e uso da informação. O processo de geração de informações ambientais é caro, levando as instituições a restringirem seu acesso e/ou uso através de citações, pagamentos ou outras formas de segurança.

Todos esses aspectos, somados a falta do poder de representação da estrutura dos dados nos padrões de metadados para SIAs, geraram a necessidade da criação de uma estrutura adequada para o armazenamento da informação estrutural dos dados ambientais para cada repositório de dados, de forma clara, mantendo-os armazenados em seus locais de origem. Para atingir tal objetivo foi criado um meta-modelo capaz de mapear a estrutura dos dados ambientais em um nível intermediário, permitindo aos usuários desenvolverem programas que acessem a informação de forma direta, sem a necessidade da troca do local de armazenamento ou de mudanças na sua estrutura de armazenamento e padrão dos dados. Esse meta-modelo faz parte de um dos componentes da arquitetura proposta para a extração de dados em SIAs, também apresentada neste artigo. De forma a facilitar a criação e manutenção de instâncias da meta-modelagem desenvolvida foi criada uma ferramenta para ambiente Internet com o objetivo de facilitar o acesso a usuários em qualquer parte do mundo. Este requisito é de essencial importância num SIA, dada a sua característica natural de distribuição das informações [6].

O restante desse artigo está organizado da seguinte forma: a seção 2 apresenta uma visão geral sobre SIAs e suas principais etapas de desenvolvimento; a seção 3 define os componentes básicos de uma arquitetura para a extração de dados em SIAs, onde o papel dos metadados ocupa um lugar de destaque; a seção 4 apresenta o meta-modelo de metadados criado, tecendo alguns comentários sobre a ferramenta de apoio desenvolvida para a gerência desses metadados; finalmente, a seção 5 expõe as conclusões gerais do trabalho, com suas contribuições e sugestões para trabalhos futuros.

2. Características dos SIAs

SIA é um sistema de gerência de informações ambientais, que manipula informações tais como solo, água, ar, clima, espécies existentes no mundo etc. [3]. A figura 2.1 apresenta três componentes importantes que participam de um sistema de informação ambiental, visando um melhor entendimento de sua estrutura de processo:

- ***Objetos Ambientais:*** representam quaisquer objetos do mundo real de interesse, tais como entidades naturais (rios, lagos, oceanos, florestas, animais, etc.) e estruturas criadas pelo ser humano (casas, prédios, pontes, fábricas, etc.);
- ***Dados de Objetos Ambientais:*** são coleções de dados que agrupam objetos ambientais. Tal coleção é uma entidade abstrata que pode ser manipulada por computadores ou por tomadores de decisão;
- ***Metadados Ambientais:*** representam os metadados ambientais utilizados para referenciar um determinado dado de objeto ambiental. Cada dado de objeto ambiental possui um ou mais objetos de metadados que determinam seu conteúdo e formato.

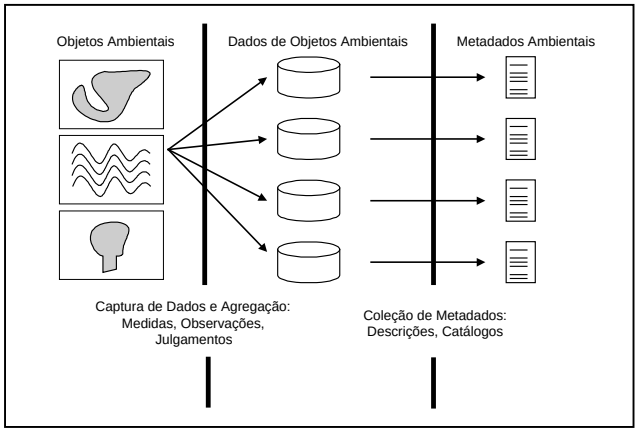

Figura 2.1 Modelo de Objetos de Três Modos de Sistemas de Informação Ambiental

Günther [3] divide, de acordo com o fluxo da informação ambiental, a estrutura de um SIA em quatro fases: *Captura de Dados, Armazenamento de Dados, Análise de Dados* e G*erenciamento de Metadados.*

A fase de *Captura de Dados* tem por objetivo a coleta, processamento e agregação dos dados ambientais em sua forma bruta (séries de medida de tempo, fotografias aéreas etc.), armazenado-os em arquivos ou em bancos de dados.

O processo da captura de dados se dá através de três etapas: processamento dos dados brutos, classificação e validação, e interpretação dos dados.

Na fase de *Armazenamento de Dados* é necessário se ter um bom projeto de banco de dados e estruturas de armazenamento apropriadas, levando a um aumento da performance do sistema. Devido a complexidade e heterogeneidade dos dados ambientais é aconselhável dispor de bancos de dados que provejam extensões à tecnologia relacional.

Na fase de *Análise de Dados* a informação disponível é preparada para propósitos de apoio à tomada de decisão.

É necessário o acesso simultâneo aos dados que estão dispersos geograficamente, armazenados em diferentes *hardwares,* e organizados em uma grande variedade de modelos de dados.

Finalmente, a fase de *Gerenciamento de Metadados* é a fase de maior importância durante o processo de criação de um SIA. Metadados são coletados e agregados nas três fases descritas anteriormente. São armazenados em estruturas de dados apropriadas, fornecendo o apoio necessário às operações de pesquisa, navegação e transferência de dados durante a fase de Análise de Dados. A criação dos metadados deve ser feita, caso seja possível, em paralelo à coleta dos dados originais, o que nem sempre ocorre.

Geralmente a criação de metadados é manual, dependendo de alguém que conheça como obter e interpretar a informação ali contida. Existem duas soluções viáveis para o problema: a primeira seria a criação de uma arquitetura organizacional que funcionasse como guia para os fornecedores dos dados durante a captura dos metadados. A segunda seria a criação de um processo de extração automática dos metadados. A primeira opção é mais voltada para as empresas privadas ou públicas que criam seus processos de extração de metadados de forma padronizada, garantindo sua qualidade. A segunda opção tem conquistado maior espaço dentre as alternativas, apesar de ainda ser difícil de ser alcançada.

Uma das soluções que facilita o processo de geração de metadados é a criação de padrões de metadados, que fornecem uma maneira sistemática de coletar os metadados. Dentre os mais importantes padrões de metadados ambientais estão o padrão americano FGDC e o europeu UDK. Ambos documentam as características dos dados ambientais, conteúdo, qualidade e condição, da mesma forma que os rótulos auxiliam o consumidor, permitindo aos usuários de tais dados determinarem se eles servem ou não para seus propósitos. Em [6] é feito um estudo detalhado desses padrões, a partir do qual foi possível concluir que: o padrão FGDC possui um maior apoio (softwares e documentação) para seu desenvolvimento e utilização pelas entidades públicas ou particulares; já o padãro UDK peca pela falta de documentação existente no idioma inglês, o que muito dificulta sua aceitação. Conforme já mencionado, a principal deficiência do padrão FGDC é não permitir representar a estrutura das fontes de dados pela seção 5, visto que essa especificação é feita em texto livre, e portanto não estruturada. A idéia seria estender essa seção, estruturando-a, de forma a abranger diversos tipos de fontes de dados.

Metadados representam portanto um módulo de extrema relevância numa arquitetura de extração dados em ambientes heterogêneos. A seguir é apresentada uma visão geral da arquitetura proposta para a extração de dados em SIAs.

3. Arquitetura de Extração de Dados em SIAs

A construção de um SIA representa uma tarefa de extrema complexidade, devido a heterogenidade da estrutura de armazenamento (planilhas, arquivos textos, banco de dados etc.) e da grande quantidade de dados disponíveis e distribuídos em diferentes locais. Tais sistemas devem permitir que os usuários consultem as informações de modo eficiente e preciso. Para isso, é necessária a união de diversas técnicas de diferentes áreas de pesquisas, tais como: metadados, ontologias, ferramentas de catalogação, classificação e pesquisa na Web, agentes etc. A integração dessas técnicas leva à definição de uma arquitetura para a extração de dados em SIAs.

A heterogeneidade dos dados ambientais e seu armazenamento em múltiplos repositórios distribuídos requer que a construção de um SIA tenha como base uma arquitetura distribuída, permitindo que seus componentes sejam integrados de forma incremental. A arquitetura proposta para a extração de dados em SIAs leva em conta essa importante característica, a partir de componentes bem determinados, permitindo que o sistema ambiental seja desenvolvido em várias camadas independentes. Além disso, pode-se utilizar uma abordagem *bottom-up*, onde o sistema é desenvolvido a partir da inclusão de segmentos independentes (fontes de dados), podendo ser integrados posteriormente, de forma distribuída, um após o outro.

Tal arquitetura deve lidar com a informação em diversos níveis de heterogeneidade, ou seja, deve trabalhar com os quatro níveis de interoperabilidade existentes [7], onde cada um é responsável por:

- **Sistema:** diferenças de hardware e de sistemas operacionais;
- **Sintaxe:** diferenças na representação dos dados, formato e armazenamento;
- **Estrutura:** diferenças de modelos de dados e estruturas;
- **Semântica:** diferenças no significado são dependentes do vocabulário e da terminologia utilizados para expressar a informação e o conteúdo na qual está interpretada.

A figura 3.1 apresenta os tipos de usuários que interagem na arquitetura, o grau de interoperabilidade e os níveis de informação [7]. Consumidores são os usuários finais que buscam a informação; fornecedores são os responsáveis pela geração dos dados (informação); e os agentes, também conhecidos como *brokers*, representam os intermediários entre os consumidores e os fornecedores, isto é, são os responsáveis pela

disponibilização da informação gerada de forma organizada pelos fornecedores, permitindo o seu acesso pelos consumidores.

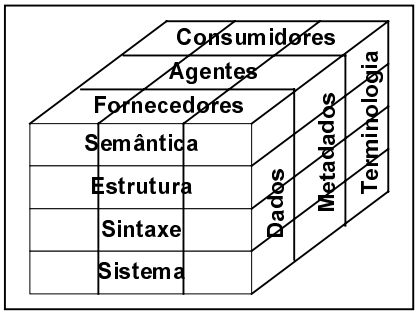

Figura 3.1 Tipos de usuários, grau de interoperabilidade e níveis de informação

Em resumo, a figura 3.1 pode ser interpretada da seguinte forma: consumidores utilizam a terminologia (ontologia) para recuperar a informação; agentes utilizam os metadados para criar processos de extração dos dados; e os fornecedores geram os dados, fornecendo-os aos consumidores.

A partir dos aspectos apresentados anteriormente é possível apresentar a arquitetura proposta para um SIA, que é composta de seis componentes, a saber: Usuário, Agente Broker, Agente Ontologia / Ontologia Específica de Domínio, Agente Extrator, Metadado e Fonte de Dados.

A figura 3.2 apresenta uma abordagem de alto nível da arquitetura proposta para SIAs dividida em 3 camadas, de forma a permitir uma melhor representação dos componentes participantes envolvidos no processo.

A seguir, é apresentada uma visão geral de cada uma das camadas:

- A primeira camada **(Usuários)** representa os consumidores de informação, ou seja, os cientistas, governantes, entidades privadas etc.;
- A segunda camada **(Aplicações)** representa os "*brokers*" ou seja, as entidades responsáveis pela integração da informação e os agentes extratores, isto é, aqueles que se comunicam diretamente com cada repositório de dados para recuperarem a informação. Cada repositório possui um ou mais extratores ou tradutores (também denominado "*wrappers*") responsáveis pela recuperação da informação solicitada. Essa segunda camada é muito importante, visto que efetua a ligação entre os consumidores de informação (Usuários) e os produtores de dados (Dados). Nela se encontram as ontologias específicas de domínio, equivalentes a um dicionário de termos, utilizadas para resolução de conflitos semânticos [7];
- A terceira camada **(Dados)** representa os repositórios de informações geradas pelos produtores de dados.

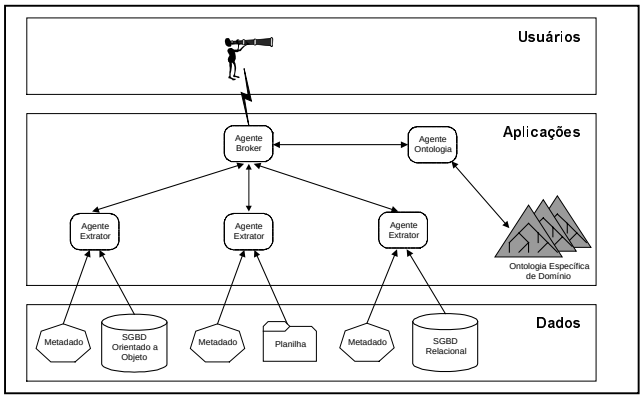

Figura 3.2 Arquitetura proposta para SIAs

A arquitetura descrita é por natureza distribuída, permitindo que se utilize a Internet como meio de transmissão de informações, levando, juntamente com a tecnologia de orientação a objetos, a um alto grau de escalabilidade e abrangência. Uma das soluções seria o uso de uma arquitetura distribuída, como a arquitetura CORBA [5] considerada atualmente como plataforma padrão para integração e distribuição de objetos. A arquitetura CORBA é uma arquitetura de objetos distribuídos, permitindo que aplicações façam solicitações aos objetos, de forma transparente, independentemente da linguagem, sistema operacional ou considerações de localização [1].

A heterogenidade das estruturas das fontes de dados (arquivos textos, planilhas, bancos de dados relacionais e orientados a objeto etc.) foi um ponto marcante para a identificação da necessidade de armazenamento de suas estruturas sob forma de metadados, além dos metadados normalmente encontrados nos padrões de catalogação de dados ambientais. O "Agente Extrator" precisa conhecer a estrutura da fonte de dados, ou seja como ela está organizada, de modo a realizar adequadamente o processo de extração dos dados. Esse conjunto de metadados, o qual denominamos de metadado estrutural, é um ponto importante no escopo desse trabalho, cuja modelagem é apresentada na seção a seguir.

4. Meta-Modelo para Suporte à Extração de Dados

Os estudos de padrões de metadados para SIAs e a definição de uma arquitetura de extração de dados em SIA levou à conclusão de que um padrão de metadados abrangente deve contemplar também metadados que representem a estrutura de armazenamento de todas as fontes de dados envolvidas no sistema. Esta seção apresenta um dos componentes necessários à construção de um SIA através de um meta-modelo denominado **M**eta-**M**odelo para **S**istemas de **I**nformações **A**mbientais (MM-SIA) [6], permitindo que os agentes extratores

recuperem a estrutura das fontes de dados a partir de um repositório central.

O MM-SIA foi desenvolvido tendo como base as seguintes fontes de dados: Arquivos Textos, com ou sem separador, Planilhas, Páginas Web (XML), Tabelas ou Visões oriundas de SGBDs Relacionais, SGBDs Relacionais-Objeto e Classes oriundas de SGBDs Orientados a Objetos. Estas fontes foram escolhidas por incluirem os tipos de fontes de dados mais comumente encontrados para o armazenamento de dados ambientais. O modelo é plenamente extensível, não impedindo que outros tipos de fontes sejam acomodados, de acordo com as necessidades dos usuários.

A idéia dessa modelagem é permitir a criação de um repositório de informações sobre a estrutura de cada fonte de dados participante de um SIA, permitindo que um agente extrator possa utilizar esse repositório para recuperar a estrutura de cada fonte de dados.

A figura 4.1 apresenta o diagrama de classes da MM-SIA modelado em UML [10], cujas propriedades são auto-explicativas.

O meta-modelo pode ser entendido da seguinte forma: um *sistema* ambiental é composto de uma ou mais fontes de dados com caraterísticas estruturais diferentes (planilhas, arquivos texto, tabelas etc.). Cada fonte de dados possui um conjunto de atributos que a define, sendo que cada um desses pode ter ou não restrições. A classe correspondente à fonte de dados *arquivo* representa arquivos textos que podem ser com ou sem separador, ou seja, um arquivo texto sem separador significa que cada coluna do arquivo possui sua posição inicial e final determinada (por exemplo, nome vai da coluna 1 a 10). Já o arquivo texto com separador possui um caracter que indica a separação entre as colunas (por exemplo, o caracter * separando as colunas nome e cpf).

A classe de atributos *atributo-arquivo* representa o local onde o atributo está localizado dentro do arquivo texto onde: C = atributo no cabeçalho; R = atributo no rodapé e E = atributo no interior do arquivo. Isso é devido a necesidade de se mapear arquivos com controles em sua primeira e/ou última linha, diferindo da estrutura do resto do arquivo.

As demais fontes de dados correspondem às seguintes classes do meta-modelo:

- ***tabela e visão***: estas classes permitem considerar respectivamente objetos e visões de sistemas gerenciadores de bancos de dados;
- ***classe***: permite representar objetos de sistemas de bancos de dados orientados a objetos e relacionais-objeto.
- ***planilha***: representa planilhas eletrônicas, a exemplo do MS Excel.
- ***XML***: permite considerar páginas no formato XML.

De forma a validar o meta-modelo apresentado, foi desenvolvida uma ferramenta de gerencia de metadados para o ambiente Web, utilizando-se como estudo de caso aplicações reais desenvolvidas na Embrapa Solos - RJ [11] e no Instituto Agronômico de Campinas - SP [12], onde grande parte dos dados encontram-se em planilhas e tabelas relacionais, tais como: solo, chuva, temperatura mínima e insolação.

Essa ferramenta foi construída com base na tecnologia do padrão Web através do uso de *applets* Java [4], possuindo uma navegação simples através de seus objetos como os estados no diagrama de transição de estados. Também permite ao usuário executar as operações básicas de gerenciamento e manipulação de cada objeto do MM-SIA.

Foram criados os metadados das informações de solo, chuva, temperatura mínima e insolação no padrão FGDC, além do mapeamento da estrutura dessas fontes de dados para o meta-modelo MM-SIA. Maiores detalhes podem ser encontrados em [6].

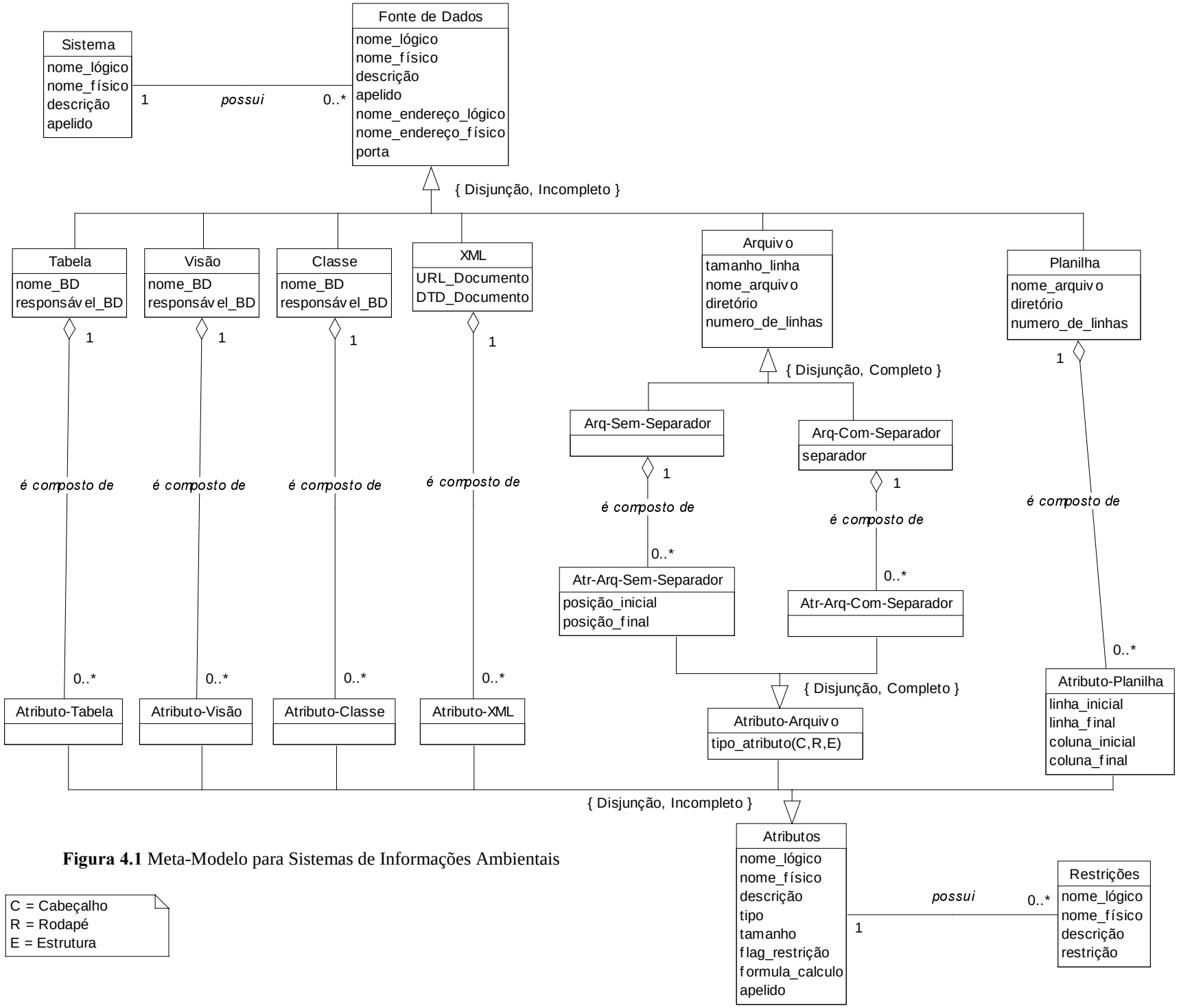

Figura 4.1 Meta-Modelo para Sistemas de Informações Ambientais

5. Conclusão

Dentre as fases de desenvolvimento de um SIA a gerência de metadados é um ponto de grande relevância. Corresponde à fase de classificação e identificação da informação ambiental, onde são definidos todos os descritores dos dados, inclusive os que descrevem a sua estrutura. A definição dessa estrutura é imprescindível durante a criação de extratores para a recuperação dos dados distribuídos, já que esses precisam conhecer a estrutura de armazenamento dos dados para poder recuperá-los de forma correta. Esta estrutura não é contemplada, por exemplo, no padrão de metadados FGDC tão utilizado pelos SIAs. Este artigo apresentou uma arquitetura genérica para a extração de dados ambientais distribuídos e um meta-modelo para a definição da estrutura dessas fontes de dados, permitindo aos donos dos dados informarem a estrutura de suas fontes de dados. Com base nesse meta-modelo foi construída uma ferramenta para gerência de metadados ambientais, tendo sido aplicada a sistemas reais de meio-ambiente. A meta-modelagem desenvolvida é extensível, podendo ser estendida a outras estruturas de armazenamento diferentes das padrões.

Como continuação desse trabalho pretende-se incluir na seção 5 (Entidade e Atributos) do padrão FGDC o meta-modelo criado, utilizando-se como fonte as instâncias já armazenadas na ferramenta desenvolvida. Para isso é necessária a extensão dessa seção de modo a contemplar os novos tipos de fontes de dados.

Referências

[1] Cattel, R. G. G. The Object Databases Standard: ODMG 2.0, Morgan Kaufman, 1997.

[2] Federal Geographic Data Committee (FGDC). http://www.fgdc.gov 1999

[3] Gunther, O. Environment Informations Systems. Springer-Verlag Berlin Heidelberg, 1998.

[4] JAVA, SUN Microsystems. http://java.sun.com/, 1999.

[5] Object Management Group http://www.omg.org, 1999

[6] Perez, H. A. M. Modelagem de Metadados para Suporte a Extração de Dados em Sistemas de Informações Ambientais. Tese de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, fev. 2000.

[7] Sheth, A., Kashyap, V., Lima, T. Semantic Information brokering: How can a Multi-Agent Approach Help?. Proceedings of the Third International Workshop on Cooperative Information Agents, julho 1999.

[8] Simon, E., Tomasic, A. Improving Access to Environment Data using Context Information, SIGMOD Record; 1997.

[9] Especificação do Modelo UDK (UmwelDatemKatalog). http://www.mu.niedersachsen.de/udk, 1999.

[10] Object Management Group http://www.omg.org/uml/, 1999

[11] Empresa Brasileira de Pesquisa Agropecuária - Centro Nacional de Pesquisa de Solos, http://www.cnps.embrapa.br, 1999

[12] Instituto Agronômico, http://www.iac.br, 1999