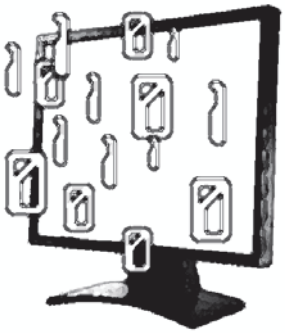
Rolf A. de By
Debora P. Drucker
José L. Campos dos Santos



Base de dados para inventários de biodiversidade

Neste capítulo, descrevemos o desenho e a implementação de um sistema para armazenar dados coletados em inventários biológicos. Esses inventários normalmente são realizados com objetivos de coletar informações sobre um grupo biológico específico em uma área geográfica. A Reserva Ducke foi o primeiro local onde os inventários do sistema RAPELD e do Programa PPBio foram conduzidos. Nesse Programa, dois tipos de levantamentos vêm sendo realizados. O primeiro trata do objetivo do PPBio de inventariar e mapear a biodiversidade, e assim pesquisar diferentes formas de vida. O segundo é o mapeamento das características geofísicas da área, realizado de forma à complementar o primeiro. Essa caracterização ambiental pode incluir, por exemplo, dados topográficos (p.ex.: altitude, declividade), composição do solo e nível da água subterrânea, além de dados climáticos (p.ex.: precipitação, temperatura e radiação).

Além de estudar a biodiversidade, o PPBio tem como objetivo tornar os dados desses levantamentos disponíveis para diferentes setores da sociedade, como comunidades locais, profissionais envolvidos com manejo (p.ex.: de fauna silvestre, de pesca, de bacias hidrográficas e florestal), laboratórios farmacêuticos, cientistas, conservacionistas e políticos.

Dados de inventários são uma importante fonte da informação na qual a pesquisa em biodiversidade é fundada. No passado, esses dados eram armazenados em cadernos de campo ou, quando computadorizados, em arquivos de texto ou planilhas. Esse tipo de arquivamento de bases de dados oferece diversas desvantagens. Primeiro, não permite que análises posteriores sejam feitas facilmente e, segundo, não garante uma sobrevida prolongada dos dados. Freqüentemente, depois da publicação de uma análise inicial, usualmente esses dados são perdidos ou esquecidos.

De qualquer forma, é comum que dados de inventários sejam mantidos em um sistema de gerenciamento de base de dados, que são conjuntos de dados armazenados com uma estrutura regular que organiza essa informação. Um tipo de base de dados muito utilizado para armazenar informações biológicas é o sistema de dados relacionais. O modelo relacional é baseado no princípio que todos os dados estão guardados em tabelas e que cada tabela é ligada a outra por uma identificação, muitas vezes denominada de chave, que nunca se repete. Porém, mais um repositório de base de dados relacional padrão não contribui significativamente para o avanço de sistemas de gerenciamento de bases de dados de inventários biológicos. Algumas características de nossa abordagem, no entanto, são inovadoras e asseguram a documentação de dados de inventários biológicos. Elas compreendem três aspectos:

1. Nosso repositório é habilitado espacialmente, o que significa que qualquer tipo de observação de inventários é georreferenciada, permitindo-nos mapeá-las e usá-las em análises espaciais ou geoestatísticas.

2. Nosso repositório foi desenhado de forma a acomodar dados de qualquer tipo de inventário, até mesmo levantamentos futuros para os quais o delineamento ainda terá que ser estabelecido. O repositório é verdadeiramente genérico, contanto que os inventários futuros sigam os princípios do delineamento geral do PPBio. Conseqüentemente, nosso repositório permite a análise futura de dados, mesmo de maneiras ainda não previstas.

3. Nosso sistema foi desenhado e está atualmente sendo desenvolvido pelos princípios de arquitetura orientada a serviços (SOA: "service-oriented architecture"). SOA é uma abordagem para o estabelecimento de aplicativos de sistema na qual as funções são organizadas como serviços, tanto para acomodar usuários externos como para organizar a colaboração entre os componentes do sistema. Isso implica em um arranjo estritamente modular que acomoda facilmente a contribuição de usuários externos, como os pesquisadores que conduzem os inventários ou os futuros usuários dos dados, enquanto permite que funções ainda desconhecidas sejam incorporadas ao sistema no futuro.

Nas próximas seções, discutiremos os requisitos para nosso repositório de inventários, seu desenho conceitual, o subseqüente trabalho de implementação, bem como os itens que deverão ser considerados no desenvolvimento futuro do sistema.

Requisitos de um repositório para dados de inventários

O principal objetivo do Sistema de Informação para Inventários Biológicos ("Biological Inventory Database" – BID) é o armazenamento e disponibi-

lização dos dados coletados em levantamentos padronizados. Apesar do processo de aquisição de dados ser padronizado de uma perspectiva biológica e estatística, não se sabe quais as características dos dados que serão obtidos e, dessa forma, tabelas definidas estaticamente não satisfazem as necessidades de gerenciamento de dados do sistema. Essas tabelas são eficientes para armazenar dados coletados em levantamentos de variáveis previamente estabelecidas, mas não para dados de futuros levantamentos de biodiversidade para os quais as variáveis ainda serão estabelecidas.

Levantamentos podem diferir bastante, o que leva a uma grande diversidade de dados ao longo do tempo. Nosso repositório deve apenas impor algumas regras fundamentais para séries de dados de levantamentos, enquanto permite flexibilidade para acomodar a maioria, senão todos, os levantamentos futuros. Um primeiro problema técnico está caracterizado aqui: **criar uma estrutura de dados verdadeiramente genérica que acomode bases de dados de todos os levantamentos.**

Muitos levantamentos de dados biológicos conduzidos com técnicas padronizadas dependem da coleta de materiais de referência para permitir estudos posteriores em laboratório. Em muitos casos, a identificação é realizada somente nessa fase: o material coletado receberá um nome, ou seja, será cientificamente identificado, bem mais tarde. Enquanto não é identificado, o material é chamado de "amostra". O sistema BID não foi desenvolvido para embasar totalmente esse trabalho futuro em laboratório, mas deve ser possível integrá-lo com os sistemas de informação de coleções que fornecerão dados para auxiliar as identificações realizadas posteriormente, depois de meses, ou até mesmo anos. **Amostras, ao longo do tempo, são divididas em unidades menores, e esta história de divisões deve ser rastreável**. Este é o segundo desafio técnico.

Quando as identificações são realizadas, deve haver um sistema de referência no qual estão baseadas. A Biologia Sistemática (Taxonomia) é a área do conhecimento responsável por fornecer tais sistemas de referência. No entanto, estes não são estáticos, pois a ciência que estuda a evolução realiza descobertas que são refletidas em alterações nos mesmos. Um sistema de informação para inventários biológicos (BID) não precisa incorporar um subsistema de dados taxonômicos, e sim contar com os vários sistemas de informação em coleções já existentes. Por exemplo, alguns desses sistemas já estão em uso no INPA: BRAHMS (sistema de banco de dados para herbários) e SPECIFY. O terceiro problema técnico é: **desenvolver o sistema de informação para inventários biológicos BID de modo que mudanças em dados taxonômicos (refletidas em alterações em dados taxonômicos nos sistemas de informação de coleções) são acomodadas automaticamente, sem manutenção posterior dentro do sistema BID.**

Desenho conceitual e arquitetura do repositório

Nesta seção, discutimos a arquitetura de aplicativos do sistema BID e providenciamos um esboço de seu conteúdo conceitual em informação. Devido a limitações de espaço, nossa discussão apenas sumariza o desenho do sistema e destaca suas características principais.

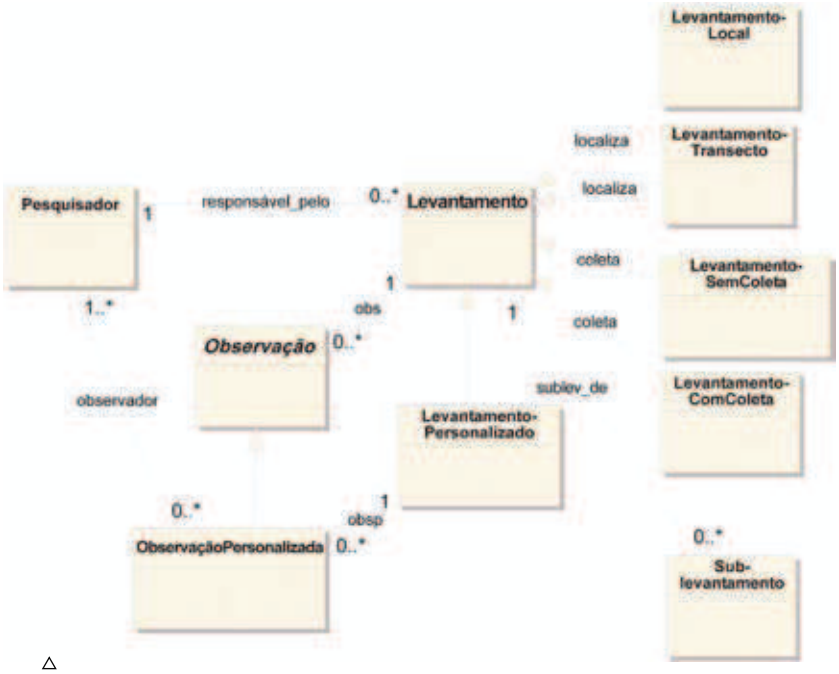
O desenho do sistema é fundado em princípios de tecnologia cliente-servidor: o repositório foi construído em um servidor de gerenciamento de banco de dados objeto-relacional, que propicia comunicação com um grande número de softwares de clientes. Além disso, esse servidor foi equipado com capacidade de armazenamento de dados espaciais e funções de manipulação, o que permite o mapeamento "on line" e a realização de análises espaciais. Os formatos de dados permitidos e as funções espaciais associadas são aqueles do "Open Geospatial Consortium" (OGC, em português Consórcio Geoespacial Aberto). Essa organização foi desenhada especificamente para servir o usuário final do sistema via Internet, ao mesmo tempo em que permite o fácil acesso de softwares dedicados de clientes ao repositório, pelo uso de conectores de banco de dados conhecidos como JDBC ("Java Database Connectivity") ou ODBC ("Open Data Base Connectivity"). A Conectividade de Banco de Dados Java (JDBC) é um conjunto de interfaces de programação padronizadas, escritas em Java, que permite a conexão com um banco de dados e o uso de suas transações SQL ("Structured Query Language"). Atualmente, a linguagem Java está muito difundida, as pessoas utilizam aplicativos em Java quando acessam contas bancárias ou assistem a vídeos na Internet, por exemplo. Por usar linguagem Java e outros conectores difundidos, nosso repositório prevê uma conectividade ampla e fácil ao maior número de aplicativos pessoais possível.

Na discussão a seguir, sumarizamos o desenho conceitual do banco de dados com alguns diagramas de classes UML ("Unified Modeling Language").

Pesquisadores e levantamentos

Nossa abordagem para o desenho do banco de dados visa buscar simplicidade e generalidade ao mesmo tempo. Fornecemos um sumário das classes de objeto mais importantes abaixo. Deixaremos de fora uma discussão detalhada de seus atributos, domínios de dados e possíveis limitações. Sugerimos que o leitor acompanhe a discussão a seguir com referência na figura 1, que é um diagrama parcial, feito para indicar somente as classes de objetos mais proeminentes.

Um **pesquisador** é um ser humano, sobre o qual registraremos certos detalhes. No presente, ele/ela podem ter dois papéis: **responsável pelo levantamento** e/ou **observador** em um levantamento. O responsável



△

Figura 1 :: Diagrama de classes UML persistente do esquema conceitual da base de dados, envolvendo levantamentos e observações.

pelo levantamento providenciará os detalhes sobre o levantamento em tempos distintos:

- antes da excursão de campo: o delineamento dos dados que serão levantados e respectivos metadados (i.e., referência aos métodos utilizados nas coletas);
- depois da realização dos levantamentos: os dados coletados e os metadados relacionados.

Um **levantamento** engloba o trabalho de campo em uma área, normalmente em localidades múltiplas dentro de uma área ao longo de um período de tempo. Observações padronizadas são realizadas em cada localidade. Essas localidades podem, por exemplo, ser posições ao longo de uma trilha, ou ao longo do curso de água.

Existem diferentes tipos de levantamentos e, conseqüentemente, as observações relacionadas a cada um deles também diferem. Levantamentos podem ou não envolver coleta de material nas localidades visitadas. Em levantamentos com coleta, às vezes os materiais são identificados no campo (como censo de aves e primatas), outras vezes são identificados posteriormente no laboratório (inventários de invertebrados e vegetação). No segundo caso, o

material coletado em uma localidade é chamado de **amostra** e registrado com um número de identificação. As amostras representam um problema específico de gerenciamento de dados que deve ser tratado de forma combinada com os sistemas de informação relevantes para coleções biológicas, como já discutimos anteriormente. Para esse fim, dado o número original da amostra, o sistema de informação em coleções utilizado deve ser capaz de gerar duas listas: **uma que contenha as informações das amostras resultantes de divisões da amostra original,** e outra lista **com todos os espécimes em coleções derivados da amostra original até o presente.** Levantamentos com coleta podem gerar observações adicionais *in situ*, como, por exemplo, descrição das circunstâncias do evento de coleta como horário ou temperatura da coleta.

Um **levantamento sem coleta** tipicamente gera um número de observações para cada local de coleta. Pode ser que estas observações originais necessitem de pós-processamento no laboratório, mas assumiremos que isso acontecerá antes dos dados entrarem em nosso sistema BID. Em outras palavras, não é necessário rastrear a história de levantamentos sem coleta.

Outra distinção entre levantamentos é se as localidades são ou não conhecidas *a priori*. Alguns levantamentos, pelo seu delineamento, listam previamente os locais exatos nos quais as observações serão feitas, denominados de **levantamentos localizados**. Outros levantamentos não definem a localização exata, mas normalmente identificam um transecto ao longo do qual as observações serão feitas, e cada observação será individualmente registrada posteriormente junto com a informação sobre o posicionamento georreferenciado, como distância ao longo da trilha e distância da trilha (à direita ou à esquerda). Chamamos esses levantamentos de **levantamentos em transectos**.

Para alguns levantamentos, é relevante registrar quem é o responsável por cada uma das observações ou eventos de coleta. Isso não é o mesmo do que registar o líder do levantamento, que indica a pessoa responsável por todo o levantamento (isso é algo que sempre deverá ser feito). Aqui nos referimos à indicação de uma pessoa para cada observação, o observador, e chamaremos um levantamento como esse de **levantamento personalizado**. Qualquer observador mencionado aqui deverá estar incluído na lista de pesquisadores.

Ainda temos que fazer mais uma distinção entre levantamentos. Isso acontecerá quando, para embasar o objetivo principal de algum levantamento, medidas ou observações adicionais que não podem ser diretamente atribuídas ou conectadas com as observações principais são feitas durante o período do levantamento. Seguem alguns exemplos:

- Em alguns estudos de herpetofauna, a temperatura do ar é importante para a coleta, e isso pode ser medido três vezes durante a coleta matinal, o que pode acontecer em mais de uma localidade. Para cada uma das medidas, teremos um valor de temperatura do ar e, possivelmente, informações de localidade e de hora.
- Em estudos de avifauna, o horário do dia de um levantamento em transecto é importante, então o horário inicial e final do levantamento devem ser registrados.

Em ambos os casos, o processamento de dados após o trabalho de campo pode produzir valores (temperatura, hora) que podem ser associados com os dados primários de observação. Mas nem sempre isso pode ser feito e, nessas circunstâncias, será permitido que medidas/observações adicionais sejam armazenadas em **sub-levantamentos** separados, relacionados com o levantamento principal. Um levantamento principal pode estar associado a mais de um sub-levantamento.

Qualquer sub-levantamento previsto requer o mesmo trabalho "a priori" que um levantamento principal (i.e., preparação de dados). Mas um sub-levantamento terá alguns aspectos de seu delineamento em comum com o levantamento principal, então é razoável atentar para eles antes da entrada de dados e metadados que pode ser de todas as classes de levantamentos acima discutidas: com ou sem coleta, localizado ou em transecto, personalizado ou não.

Observações e variáveis

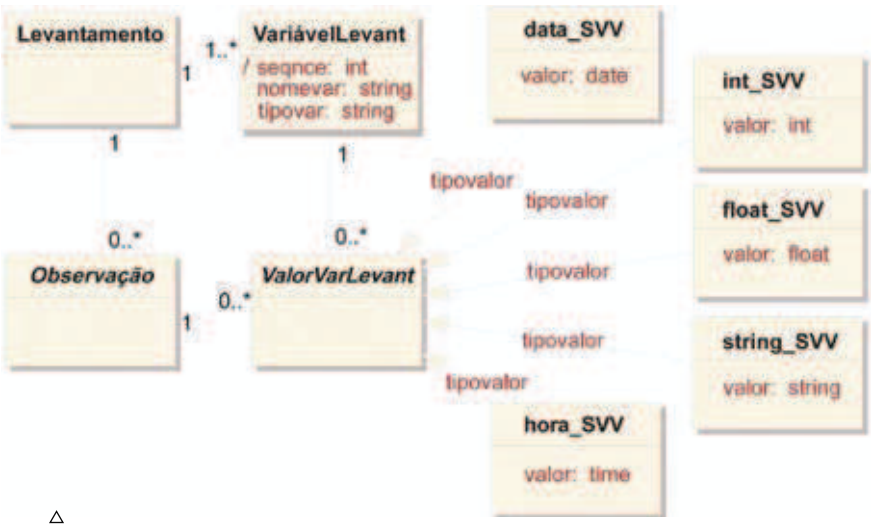
Qualquer tipo de levantamento envolve observações, e os dados então registrados são inseridos no sistema BID. Para ajudar a compreensão podemos pensar nos dados registrados em um levantamento como uma tabela única, na qual cada linha representa todos os dados coletados em uma só localidade do levantamento (em um determinado momento). Chamamos essa combinação de dados em um único local em um momento de observação. A classe de objeto para observações aparece na figura 1 e é repetida na figura 2. Seu nome em itálico indica que se trata de uma classe abstrata: não pode ser diretamente instanciada. As instâncias procedem de suas subclasses.

Como as observações são tão diferentes, criamos uma lista de subclasses de Observação. São elas: **ObservaçãoPersonalizada**, **ObservaçãoComColeta**, **ObservaçãoComData** e **ObservaçãoComDataEHora**, bem como quatro tipos de classes de observação dependendo da localidade, de acordo com as várias localidades que serão mencionadas e explicadas em detalhe na figura 4. As propriedades dessas subclasses podem até mesmo ser combinadas de forma que, por exemplo, uma observação com coleta possa ser datada e com hora.

Uma coluna nessa tabela imaginária é o que chamamos de **variável do levantamento.** Um dos objetivos principais do delineamento de um levantamento é a identificação das variáveis do levantamento. Isso determina quais **valores de observação** serão gerados em cada localidade do levantamento. Novamente, a classe Valor da Variável Levantada (**ValorVarLevant**) é abstrata. Possui cinco subclasses (Figura 2), todos os nomes terminando com **SVV**. Cada uma dessas classes tem um valor extra de atributo, no qual o tipo de dado varia por classe. Essa técnica está no cerne da solução para obter a generalidade do inventário.

Vamos discutir a técnica descrita acima com um exemplo. Em um levantamento, algumas medidas serão realizadas. Suponhamos que estamos fazendo um levantamento em parcelas aquáticas, e que para cada parcela determinamos o total de sólidos dissolvidos ("TDS", em inglês) em medidas de [mg/l]. No sistema BID, essas medidas criarão um objeto único do **Levantamento**, um objeto da **VariávelLevantada** com um nome **TDS** (varname) e um tipo de variável tipo "**float**". Esses objetos da **VariávelLevantada** estarão associados com tantos objetos **float_SVV** quanto parcelas aquáticas visitadas no levantamento. Obviamente, haverá mais variáveis levantadas, cada uma delas associada com seus respectivos valores.

Além do nome da variável levantada, no delineamento do levantamento temos também que identificar seu tipo de dado (alfa-numérico, número integral, número real ou outro), decidir se permitiremos entradas de valores



△

Figura 2 :: Diagrama de classes UML persistente do esquema conceitual da base de dados, envolvendo variáveis levantadas, observações e valores observados.

nulos[1] para a variável e a possibilidade de valores especiais: algumas vezes valores ordinários são atribuídos com significados especiais. Por exemplo, o valor "0.00" pode significar "valor baixo demais para ser medido". Valores especiais devem sempre ser registrados, pois caso contrário sua interpretação correta pode estar em risco. Uma tabela separada será usada para administrar valores especiais de variáveis levantadas.

Algumas variáveis levantadas requerem tratamento especial pelo sistema, e isso deve ser indicado pelo responsável pelo delineamento do levantamento. Temos que saber qual combinação de variáveis levantadas pode ser usado como um identificador de uma observação (i.e., a noção de chave primária). Para um levantamento com uma única observação por parcela, o sufixo identificador da parcela; para aqueles com levantamentos múltiplos por parcela, variáveis adicionais levantadas (que podem ser localização, hora e/ou colunas do observador) são necessárias para identificar aquela observação como única. O motivo pelo qual isso deve ser conhecido tem dois aspectos:

1 - Trata-se de uma decisão importante no delineamento do levantamento, então não deve ficar sem resposta;

2 - Nossa técnica para obter generalidade no manuseio de um arquivo de dados de um levantamento conta com uma radical fragmentação vertical dos arquivos de dados. Para fazer isso, precisamos entender o que constitui a chave primária da tabela original. Se a tabela original tem N atributos nos quais P são atributos que compõem a chave primária, a fragmentação vertical levará, pelo menos ao nível de lógica de banco de dados, a N – P novas tabelas, cada uma com P + 1 atributos, nominalmente os atributos originais da chave primária mais um atributo normal de dados.

Para as variáveis levantadas, identificamos um número de subclasses, coincidindo com os vários tipos de dados que valores de observações possam ter. Isso leva a uma fragmentação vertical da tabela de dados imaginária, a qual nos permite tecnicamente a obtenção da generalidade e da independência do delineamento do levantamento.

Razões adicionais para o tratamento especial de alguns atributos do levantamento estão na demanda da verificação de sua consistência lógica pelo sistema. Em outras palavras: localidades identificadas nos dados devem existir, ou pelo menos ser representações razoáveis, observadores devem ser conhecidos pelo sistema, dados de data e hora devem estar no intervalo de tempo do levantamento e assim por diante. Também é previsível que alguns atributos do levantamento sairão de uma lista de valores possíveis, e queremos verificar a consistência desse tipo de dados.

1 Um valor nulo representa o fato de que nenhum valor foi obtido, que o valor foi perdido ou que não pôde ser obtido. Em geral, significa "valor desconhecido ou não aplicável".

Localidades de Levantamentos

O sistema BID é um sistema de banco de dados habilitado espacialmente, o que nos permite rastrear as localidades dos levantamentos, reservas e trilhas, bem como agregar espacialmente dados de levantamentos em unidades maiores, como sistemas completos de cursos de água, bacias hidrográficas ou interflúvios. No delineamento do sistema, nós consideramos uma quantidade de formas espaciais comumente encontradas em dados de levantamentos. Muito da discussão dessa seção está ilustrada na figura 3.

Em nossa modelagem espacial, utilizamos os padrões amplamente conhecidos, como ponto e polígono, e funções relacionadas de acordo com o estabelecido pelo OGC ("Open Geospatial Consortium"). Esse padrão nos permite armazenar e manipular formas espaciais de modo que podemos utilizá-las para mapeamento, análises espaciais e geoestatísticas, dentre outras.

Sobre o sistema de referência espacial

Todos os dados espaciais armazenados no sistema foram escolhidos para ser representados em um sistema de referência único, WGS 84, o qual é amplamente conhecido como o sistema padrão que a maioria dos instrumentos de posicionamento global (aparelho GPS) opera. A vantagem dessa escolha é que utilizamos um único sistema; a desvantagem é que ele não permite cálculos diretos de distância e área porque é um sistema de referência geográfica. Uma alternativa óbvia seria utilizar as várias zonas projetadas UTM, as quais contornariam essa desvantagem. A abrangência da área estudada pelo

Figura 3 :: Diagrama de classes UML do esquema conceitual do banco de dados, envolvendo reservas, trilhas e parcelas de coleta. Essas classes têm conotações de formas espaciais, indicadas por um ícone: multipolígono, linha e ponto, respectivamente.

△

Figura 4 :: Alguns objetos com conotações de formas espaciais: uma reserva (um polígono, em amarelo), um igarapé e sua cabeceira (uma linha, em verde), uma trilha e seu início (uma linha, em vermelho), e algumas parcelas aquáticas (pontos, em azul).

PPBio, entretanto, cobre pelo menos onze zonas UTM, e o uso da projeção UTM implicaria na impossibilidade de comparação direta entre alguns dados. O sistema de banco de dados incorpora funções eficientes que permitem transformar de um sistema de referência para outro, então nossa escolha pelo WGS 84 é adequada e é possível lidar com suas deficiências.

A escolha pelo WGS 84 como sistema de referência espacial para ser usada internamente não interfere em como o usuário final pode fornecer seus dados espaciais, ou como o mesmo pode obter dados espaciais pelo banco de dados. O sistema final acomodará outros sistemas de referência para entrada e saída de dados espaciais.

Áreas Primárias de Levantamento

Uma **reserva** é uma área de estudo, normalmente com um sistema de trilhas, e possivelmente abrangendo igarapés. Usamos o termo reserva no sentido amplo, relacionado a qualquer sítio de estudo no qual as pesquisas são realizadas. A representação típica de uma reserva é uma forma poligonal, ou um conjunto de formas poligonais, no caso de áreas desconexas. Uma trilha pertence a uma reserva. A representação espacial de uma trilha é uma linha com um início definido. É possível realizar medidas

de distância ao longo da trilha, pois localidades às vezes são indicadas por essas distâncias. Como uma trilha, um **igarapé** é normalmente associado a uma reserva. Assumiremos também que temos uma cabeceira de igarapé pré-determinada espacialmente (por algum método adotado).

Uma série de dados consiste em uma coleta de informações, cada uma para uma única localidade, ou possivelmente cada uma para cada combinação de localidade e hora. O nível mais grosseiro de determinação de localidade é a **parcela**. Há parcelas de pelo menos três tipos: **parcelas terrestres**, **aquáticas** e **ripárias**. Cada uma dessas normalmente é identificada com um nome que é formado pela concatenação do nome alfa numérico da trilha, mais a distância em metros do início da trilha (ver Introdução, Figura 6). No caso de parcelas ripárias ou aquáticas, usa-se às vezes o nome do igarapé. É possível derivar uma localidade a partir desses dados se a trilha e o início da trilha são geograficamente referenciados.

Do ponto de vista do gerenciamento dos dados, a distinção entre os três tipos de parcela é praticamente sem importância. A diferença principal entre os tipos de parcelas é no arranjo espacial na área de estudo: parcelas terrestres normalmente são distribuídas em uma grade regular com distâncias fixas pré-estabelecidas; parcelas aquáticas são posicionadas com distâncias normalmente determinadas pelo cruzamento entre as trilhas e os cursos de água, assim como as parcelas ripárias, estas localizadas adjacentes aos igarapés.

Outras Localidades de Levantamento

Alguns levantamentos requerem a determinação da localização em uma escala mais fina do que parcela. O termo que aplicamos para esses locais é **localidade**. Uma **localidade baseada na trilha** é uma localidade determinada pelo nome da trilha e uma distância ao longo da mesma. É muito parecida com as parcelas terrestres, aquáticas e ripárias que vimos acima, mas uma diferença fundamental é que localidades são específicas de determinado levantamento: não haverá um repositório central para elas, e as mesmas serão armazenadas com a série de dados do levantamento relacionado. Dentre as localidades baseadas na trilha, reconhecemos ainda dois tipos: **localidade ao longo da trilha** e **localidade fora da trilha**. A primeira consiste em nome de trilha mais uma distância do início da trilha; a segunda adiciona a essa uma distância medida perpendicular da trilha para a localidade fora da trilha. Por convenção, começando do início da trilha, qualquer distância para a esquerda da trilha é indicada com números negativos, qualquer distância para a direita com números positivos. A unidade é sempre metros.

O trabalho de implementação

O sistema BID, para o qual discutimos as principais características de delineamento nas seções anteriores, é desenvolvido em PostgreSQL, um sistema de gerenciamento de banco de dados de código aberto de alta qualidade. Ele implementa um modelo de dados objeto-relacional e, com a extensão PostGIS que usamos, torna-se um servidor de dados poderoso e habilitado espacialmente. Ele permite disponibilizar o conteúdo de dados, como os de levantamentos e dados espaciais relacionados, a uma grande variedade de clientes. O mecanismo mais importante de disponibilização ao usuário por enquanto é o Portal do PPBio na Internet (http://ppbio.inpa.gov.br/), no qual pesquisadores poderão fazer a entrada e saída das séries de dados de inventários.

O trabalho de implementação se concentrou até agora especialmente no desenvolvimento das estruturas de armazenamento principais, que são tabelas derivadas metodicamente do desenho conceitual do sistema, conforme ilustrado nas figuras 1, 2 e 3. Esses diagramas descrevem o sistema apenas parcialmente, e não revelam a estrutura completa da base de dados. Apresentam, porém, a parte mais importante do sistema.

A generalidade obtida por um desenho cuidadoso de alto nível precisa ser funcionalmente implementada, e isso é a parte tecnicamente desafiadora do sistema BID. Colocamos bastante ênfase na funcionalidade primária I/O ("Input/Output", ou Entrada/Saída), baseados em nosso entendimento do ciclo de informação associado a um levantamento.

Os passos principais do ciclo são:

- Delinear o levantamento e torná-lo conhecido pelo sistema BID;
- Executar o levantamento e coletar os dados em fichas de campo ou equipamentos de armazenamento digital (*dataloggers*, *palm tops*);
- Preparar os dados em arquivos .csv ("comma separate values", valores separados por vírgula) e inseri-los no sistema BID (*upload*);
- Fragmentar os dados inseridos e reorganizá-los nas tabelas internas apropriadas do sistema BID.

Um ciclo de informação semelhante foi definido para pessoas que desejam extrair dados de levantamentos armazenados no sistema BID, mas não discutimos tal processo nesse capítulo.

A complexidade de fornecer a funcionalidade de generalidade é ilustrada com códigos para o procedimento "**bid_grind_strobs**" que incluímos

aqui. Esse procedimento, implementado na linguagem pgpsql do PostgreSQL, fragmenta um arquivo de dados entrado no sistema em pedaços e armazena esses pedaços nas respectivas tabelas do banco de dados. Esse código contém comentários compreensíveis para permitir que o leitor não especializado em Tecnologia da Informação (TI) tenha uma idéia do objetivo de seu conteúdo. Como não temos a intenção de fazer um exercício de compreensão de códigos, explicamos que este código ilustra duas características proeminentes de nossa abordagem para a implementação.

A primeira é a de sumarização: as primeiras 15 linhas do código garantem que os dados de entrada são apropriados e que não aceitamos dados com defeitos ou inconsistentes com o sistema BID. A segunda característica é a da generalidade: o procedimento analisa o tipo de levantamento e trabalha de acordo, incluindo atributos relevantes, dependendo do tipo de levantamento. Dessa forma, utiliza o procedimento interno "**bid_grind_strobs**", que realiza o armazenamento de valores de variáveis do levantamento nas várias tabelas SVV que já discutimos. Outros procedimentos armazenados no sistema atual mostram itens similares de sumarização e generalidade.

Tabela 1 :: Código para o procedimento "bid_grind_strobs".

```
 CREATE OR REPLACE FUNCTION
bid_grind_strobs(levantnr integer)
  RETURNS character varying AS
$BODY$
  --- This function intends to create records in STROBS, or one of its subtables,
  --- one for each record in levantnr's tmpdatatable, which is the imported csv file.
  --- Upon import, famfresh strid< key values have already been created, and so
  --- these can be used for insertion in strobs also.
DECLARE
  rec RECORD;
  tablename varchar;
  cnt int4;
  attlist varchar;
BEGIN
  --- Verify whether survey with given number levantnr exists, if not, bail out.
  IF NOT EXISTS (SELECT * FROM levantamento AS l WHERE l.lidx=levantnr) THEN
    RAISE EXCEPTION 'Levantamento with identifier number % does not exist.', levantnr;
  END IF;
  --- See whether it has a temporary data table, if not bail out
  SELECT l.tmptablename INTO tablename FROM levantamento AS l WHERE l.lidx=levantnr;
  IF tablename IS NULL OR length(tablename)=0 THEN
    RAISE EXCEPTION 'No temporary data table seems to be currently associated with levantamento %.', levantnr;
  END IF;
  --- Verify whether tablename exists, if not, bail out.
  SELECT * INTO rec FROM pg_catalog.pg_class as c WHERE c.relname=tablename;
  IF NOT FOUND THEN
    RAISE EXCEPTION 'No table % exists in this database.', tablename;
  END IF;
  -- Verify whether the strobs of this survey have not already been inserted, if they have been, bail out.
  SELECT count(*) INTO cnt FROM strobs WHERE levant = levantnr ;
  IF cnt > 0 THEN
    RAISE EXCEPTION 'A number of % strings-of-observation have already been entered into this database for the given survey.', cnt;
  ELSE
    --- Let's meet our intentions, and insert the strobs in one or more of the strobs tables.
    --- We need to properly analyse which subtable(s) require filling.
    --- We go one-on-one.
    SELECT * INTO rec FROM levantamento AS l WHERE l.lidx=levantnr;
    --- First the positioning subtypes:
    IF      rec.locational_kind='g'  THEN  attlist = 'x,y';
      ELSIF rec.locational_kind='p'  THEN  attlist = 'parcela';
      ELSIF rec.locational_kind='t1' THEN  attlist = 'trilha, dist';
      ELSIF rec.locational_kind='t2' THEN  attlist = 'trilha, dist, dist_off';
      ELSIF rec.locational_kind='i1' THEN  attlist = 'igarape, dist';
      ELSIF rec.locational_kind='i2' THEN  attlist = 'igarape, dist, dist_off';
    END IF;
    PERFORM bid__insert_strobs(rec.tmpdatatable,rec.locational_kind,attlist);
    --- Second, those deriving from collects
    IF rec.is_collecting_survey THEN
      PERFORM bid__insert_strobs(rec.tmpdatatable,'c','lote');
    END IF;
    --- Third, those deriving from timkind
    IF rec.timing_kind='d' THEN
      PERFORM bid__insert_strobs(rec.tmpdatatable,'d','date');
    ELSIF rec.timing_kind='h' THEN
      PERFORM bid__insert_strobs(rec.tmpdatatable,'h','datum, hora');
    END IF;
    --- Fourth, those deriving from personal
    IF rec.is_personalized_survey THEN
      PERFORM bid__insert_strobs(rec.tmpdatatable,'p','obs_pesq');
    END IF;
    RETURN 'Strings-of-observation properly included in strobs table.';
  END IF;
END; $BODY$
  LANGUAGE 'plpgsql' VOLATILE;
ALTER FUNCTION bid_grind_strobs(integer) OWNER TO postgres;
```

Desenvolvimento futuro

O trabalho já conduzido no desenho e implementação do sistema focou na construção de um sistema de armazenamento e de uma interface na internet para permitir o registro do delineamento de levantamentos, a entrada de dados observados e a saída de dados de levantamentos já conduzidos. Todas as funções estão prontas e o sistema está em fase de teste.

Para uma segunda fase desse projeto, é necessário o estudo de requerimentos para o mapeamento mais avançado e para análises estatísticas e geoestatísticas. Isso nos permitirá construir ferramentas de exploração de dados coletados em uma variedade de formas diferentes. Alguém pode pensar em estudos integrados em reservas ao longo do tempo e de vários grupos de formas de vida, ou em estudos de um só grupo de organismo ao longo de várias reservas. O sistema, quando estiver totalmente constituído, atenderá essas demandas.

No futuro próximo, adicionaremos ao sistema os levantamentos já conduzidos na Reserva Ducke. E, após ajustes e adequações, adicionaremos levantamentos conduzidos em outras áreas de estudo com o delineamento RAPELD, como Parque Nacional do Viruá, Estação Ecológica de Maracá, Universidade Federal de Roraima, Embrapa de Boa Vista, Reserva Biológica do Uatumã, módulos de coleta ao longo da BR-319 e Floresta Nacional de Caxiuanã. Isso nos fornecerá dados que poderão ser experimentados para o desenvolvimento de formas avançadas de mapeamento de ferramentas geoestatísticas. Essas ferramentas também serão disponibilizadas no Portal PPBio na Internet.

Será realizado também mais trabalho na distribuição dos dados levantados em formatos apropriados de metadados. Estamos atualmente investigando qual(is) formato(s) são mais indicados para isso. Nosso objetivo é facilitar o acesso aos dados armazenados pela maior gama de usuários finais possível.

Agradecimentos

Agradecemos Célio Magalhães e o INPA por viabilizar a estadia de Rolf de By e o ITC (International Institute for Geo-Information Science and Earth Observation) por permitir o afastamento sabático de Rolf. Agradecemos também o apoio de Bill Magnusson ao longo de todo o processo de desenvolvimento do sistema e redação desse capítulo e Luciano Ferreira e colaboradores do Projeto LBA na implementação do sistema.

Sugestões de leitura

Geoffrey C. Bowker. Biodiversity datadiversity. Social Studies of Science, 30(5):643–683, October 2000.

Jim Arlow and Ila Neustadt. UML 2 and the Unified Process: Practical Object-oriented Analysis and Design. Addison-Wesley, second edition, 2005.

OpenGIS Consortium. OpenGIS simple features specification for SQL. Technical Report 99–049, OpenGIS Consortium, Inc., May 1999. revision 1.1.

PostgreSQL Global Development Group. PostgreSQL 8.2.0. Documentation. Technical report, The PostgreSQL Global Development Group, 2006. 1688 pp.

William K. Michener. Meta-information concepts for ecological data management. Ecological Informatics, 1(1):3–7, 2006.