Sistema de Informação Distribuído para Coleções Biológicas:

a Integração do *Species Analyst* e *Sin*Biota

Coordenador: Vanderlei Perez Canhos

Processo FAPESP: 2001/02175-5

Índice

“Sistema de Informação Distribuído para Coleções Biológicas: a Integração do *Species Analyst* e *Sin*Biota”

Coordenador: Vanderlei Perez Canhos

Processo FAPESP: 2001/02175-5

1. RESUMO

O projeto tem por objetivo estruturar um sistema distribuído de informação com dados dos acervos das Coleções Biológicas do Estado de São Paulo. A meta é criar mecanismos tecnologicamente viáveis que permitam a recuperação dos dados dos acervos em tempo real através de um portal padrão, mantendo, porém, o domínio dos respectivos curadores sobre a disponibilização e atualização dos dados.

O projeto também tem como meta técnica respeitar a liberdade de cada coleção quanto à escolha do *software* utilizado para o gerenciamento de seu acervo. A idéia básica é que o sistema de recuperação de dados interfira o mínimo possível na filosofia, metodologia e rotina de trabalho de cada coleção, sendo um elemento o mais transparente possível tanto para a coleção provedora de dados quanto para o usuário que vem em busca desses dados.

Apesar de não ter sido um requisito do projeto, optou-se por uma arquitetura baseada em *software* livre e de protocolos abertos. O sistema está implementado em computadores Intel, sistema operacional Linux Red Hat, usando servidor web Apache, linguagens de programação Perl, PHP e Java, utilizando protocolos HTTP, SOAP e XML.

Depois de analisar várias possibilidades de protocolos que poderiam ser utilizados para a normatização de dados, optou-se pela utilização do protocolo DiGIR que está sendo desenvolvido pela equipe do CRIA em colaboração com a equipe da Universidade do Kansas (e outros parceiros) como *open source* e que tem um grande potencial de se tornar o protocolo internacional para interoperabilidade de dados biológicos.

Estudos mostraram que grande parte das coleções brasileiras tem problemas com a conectividade, infra-estrutura computacional (*hardware* e *software* para o gerenciamento dos acervos), e disponibilidade de recursos humanos habilitados a digitalizar, validar e manter os dados dos acervos. As coleções do Estado de São Paulo não são exceção. A maioria possui micro computadores bastante simples, tem acesso bastante precário à Internet e eventualmente conta com a colaboração de estagiários ou bolsistas para a digitalização dos dados. Assim, somente parte dos acervos está digitalizada e os sistemas de gerenciamento utilizados são pouco específicos ou adequados para o trabalho.

Chegou-se à conclusão que para viabilizar o projeto, além de uma infra-estrutura básica (*hardware* e *software*), seria necessário oferecer suporte às coleções na escolha e utilização de *software* adequado. Foi feito um estudo detalhado sobre os vários *software* existentes para gerenciamento de coleções biológicas, levando-se em conta fatores como estabilidade, adaptabilidade, suporte, especificidade e capacidade de importação e exportação de dados. Com cada coleção foi definido qual o melhor *software* a ser utilizado, quando necessário foi oferecido treinamento e foi dado suporte para a migração dos dados e operação do sistema. Para que pudessem participar plenamente do projeto, as coleções receberam infra-estrutura mínima (microcomputador e *software*)

e, em alguns casos, a instalação de pontos de rede e equipamentos complementares (p.ex. *no-breaks*). Além disso, alguns elementos intermediários precisaram ser introduzidos no desenho inicial do sistema de recuperação de dados para que vários problemas limitantes no esquema cliente ↔ servidor (portal ↔ provedor) pudessem ser resolvidos.

Foram então introduzidos os conceitos de **servidores locais** e **regionais** que viabilizassem o acesso em tempo real aos dados das coleções. Esses servidores têm a capacidade de manter um espelho do banco de dados das coleções. Através de um *software* especialmente desenvolvido em Java, o **spLinker**, o curador, através de um clique em um botão, pode transferir os dados atualizados da sua coleção para o servidor local ou regional, assim como, desativá-lo, se assim o desejar.

Além das coleções biológicas, o sistema species**Link** também integra os dados disponíveis no *Sin*Biota e estudos já estão sendo feitos para a integração do sistema a outros provedores de dados internacionais que também utilizam o protocolo DiGIR. Pretende-se dessa forma ter acesso a dados de qualidade em quantidade suficiente para serem utilizados em várias outras aplicações como análises estatísticas, modelagem de nicho ecológico de espécies, mapeamento de espécies, etc.

Em paralelo ao desenvolvimento do sistema de recuperação de informação, objeto central do projeto, e do apoio à informatização dos acervos foram desenvolvidas várias outras atividades visando a apresentação e a utilização dos dados. O relatório apresenta a pesquisa desenvolvida no campo da modelagem preditiva de distribuição de espécie baseada em nicho ecológico.

2. INTRODUÇÃO

O projeto tem por objetivo desenvolver um sistema distribuído de informação com os dados dos acervos de coleções biológicas do Estado de São Paulo. Pretende também desenvolver aplicativos que façam uso dessa infra-estrutura de dados.

As coleções que se comprometeram a participar desta fase do projeto são:

- Coleção Brasileira de Microrganismos de Ambiente e Indústria, CBMAI - CPQBA/UNICAMP
- Coleção de Culturas de Fitobactérias do Laboratório de Bacteriologia Vegetal, IBSBF - Instituto Biológico de Campinas
- Coleção de Fungos do Herbário do Estado "Maria Eneyda P. Kaufmann Fidalgo", SP - Instituto de Botânica, IBt
- Coleção de Algas do Herbário do Estado "Maria Eneyda P. Kaufmann Fidalgo", SP - Instituto de Botânica, IBt
- Coleção de Fanerógamas do Herbário do Estado "Maria Eneyda P. Kaufmann Fidalgo", SP - Instituto de Botânica, IBt
- Herbário da Universidade Estadual de Campinas, UEC - IB/UNICAMP
- Herbário do Departamento de Botânica, SPF - IB/USP
- Coleção de Ácaros do Departamento de Entomologia, Fitopatologia e Zoologia, AcariESALQ - LEF/ESALQ
- Coleção de Ácaros do Departamento de Zoologia e Botânica, AcariDZSJRP - IBILCE/UNESP
- Coleção de Peixes do Departamento de Zoologia e Botânica, DZSJRP - IBILCE/UNESP
- Coleção de Peixes do Laboratório de Ictiologia de Ribeirão Preto, LIRP - FFCLRP/USP
- Coleção de Peixes do Museu de Zoologia, MZUSP - IB/USP

Além da integração dos dados dos acervos das coleções, a proposta visa também a integração do sistema com o *Sin*Biota[1] e o projeto *Species Analyst*[2].

As principais atividades do projeto são:

- Estudo de *software* para gerenciamento de coleções biológicas para auxiliar as coleções na informatização de seus acervos e para a criação de interfaces com o servidor de busca;
- Interação com as coleções;
- Desenvolvimento de um protocolo cliente/servidor para recuperar informação de fontes distribuídas;
- Desenho da arquitetura da rede speciesLink baseado nas características de cada coleção (*software* utilizado, infra-estrutura disponível (*hardware* e conectividade)) e na tecnologia disponível;
- Integração com outros sistemas: *Sin*Biota, *Species Analyst* e outras redes de coleções científicas internacionais;

[1] http://sinbiota.cria.org.br
[2] http://www.speciesanalyst.net/

- Desenvolvimento de aplicativos: modelagem de distribuição de espécies;
- Desenvolvimento de um website com informações sobre o projeto e sistema de busca e recuperação de dados;
- Difusão: participação em reuniões científicas e publicações; e,
- Formação de Recursos Humanos.

3. ATIVIDADES REALIZADAS

3.1. ESTUDO DE *SOFTWARE* DE GERENCIAMENTO DE COLEÇÕES

O primeiro passo foi trabalhar com a infra-estrutura de dados, base para todo o desenvolvimento do projeto. Embora grande parte das principais coleções científicas paulistas tenha se modernizado nos últimos anos devido a incentivos, em especial providos pela Fapesp, a situação das diferentes coleções é bastante heterogênea. Encontramos coleções informatizadas, parcialmente informatizadas e outras em processo de escolha do *software* a ser utilizado.

Como se trata de um projeto de pesquisa, consideramos importante lidar com todas as situações, daí a escolha de coleções em estágios tão diversos. O único critério comum foi o compromisso de compartilhar os dados através de um sistema de acesso público na Internet.

Para auxiliar as coleções científicas no processo de informatização e para estudar formas de integrar os diferentes acervos, foi realizada uma avaliação preliminar dos *software* disponíveis no mercado para a informatização de coleções biológicas. Os *software* estudados foram:

- Biota (Robert Colwell)
- Brahms (Universidade de Oxford)
- Specify (Universidade do Kansas)
- Microsoft Access e Sistemas Gerenciadores de Bancos de Dados Relacionais
- Planilha Microsoft Excel

3.1.1. BIOTA (ROBERT COLWELL)[3]

O Biota Collections Management System é um *software* desenvolvido para gerenciar bancos de dados de coleções biológicas tanto nas áreas de botânica quanto de zoologia. Foi inicialmente desenvolvido em plataforma MacIntosh e depois portado para o ambiente Microsoft Windows.

Entre as principais características positivas do produto, podem ser listadas:

- Fácil aprendizado. O modelo de dados do Biota é composto por tabelas principais (coleta, espécimen, localidade, equipe, etc.) o que torna simples o aprendizado.
- Interface gráfica bastante intuitiva e coerente.
- Documentação. O manual de instruções é bem completo e o texto é apresentado de forma bastante didática.
- Estabilidade. Não houve nenhum caso de perda de dados ou “travamento” do sistema enquanto era usado.
- Grande mobilidade de dados. É extremamente fácil importar ou exportar dados no Biota, seja com relação às tabelas principais, ou com relação à base de dados completa. O processo de importação e exportação usa, como meio de transferência, arquivos texto com campos separados por espaços e permite a inclusão de um subconjunto dos campos do arquivo e a concatenação de campos, entre outras facilidades.

[3] http://viceroy.eeb.uconn.edu/Biota

- Facilidade em manter cópias de segurança (*backup*);
- Preço acessível para uso local por um único usuário.

As características negativas encontradas foram:

- Versão básica não funciona em rede. Para utilizar o *software* em rede num cenário de múltiplos usuários acessando a base de dados simultaneamente (cenário comum para as coleções de maior porte), é necessário adquirir a versão para servidor do gerenciador de bancos de dados usado pelo Biota, chamado 4th Dimension Server, que custa cerca de US$1.000,00, para cada 5 usuários simultâneos.
- Impossibilidade de personalização da interface para atender a casos específicos de uso.
- Falta de opções para geração de relatórios e resumo de dados mais elaborados. Força o usuário a exportar os dados e visualizar os dados em outros sistemas.

3.1.2. BRAHMS (UNIVERSIDADE DE OXFORD)[4]

O Brahms é um *software* desenvolvido especificamente para atender às necessidades de coleções botânicas. Foi desenvolvido para ambiente Microsoft Windows utilizando o modelo de dados do FoxPro.

Entre as principais características positivas do produto tem-se:

- Atende às necessidades específicas da área de botânica.
- Permite a geração de relatórios mais complexos.
- Bastante difundido e adotado pela comunidade da área de botânica no Brasil.
- Funciona em rede com múltiplos usuários acessando o sistema simultaneamente.
- É gratuito.
- Facilidade em mover os dados para dentro e para fora do sistema. Apesar de ser um pouco mais complicado que o Biota para importar os dados, o sistema permite a importação de planilhas para o banco de dados nativo. A exportação dos dados também é simples, permitindo mover os dados com facilidade.

Entre as principais características negativas do produto tem-se:

- Interface gráfica pobre. Os dados são sempre mostrados em formato de planilhas de cálculo, com um número de colunas que excede em muito o espaço disponível do monitor, obrigando o usuário a freqüentemente arrastar o cursor de um lado a outro.
- Documentação quase inexistente. Isso obriga o usuário a aprender a usar o sistema através do método de tentativa e erro, que por sua vez é dificultado pela interface gráfica pouco intuitiva.
- Modelo de dados é escondido pela interface. Embora o sistema use bancos de dados relacionais, os dados são sempre mostrados como uma grande planilha. Isso dificulta o aprendizado do modelo de dados utilizado pelo sistema e o seu funcionamento, o que leva o usuário muitas vezes a cometer erros na entrada ou importação dos dados.

[4] http://storage.plants.ox.ac.uk/brahms/

3.1.3. SPECIFY (UNIVERSIDADE DO KANSAS)[5]

O *software* Specify, desenvolvido pela Universidade do Kansas, é projetado para atender às necessidades de uma ampla variedade de coleções biológicas, desde coleções botânicas e zoológicas até coleções paleontológicas.

Como características positivas, tem-se:

- Modelo de dados completo. Contém cerca de 70 tabelas de dados que podem ser usadas de diversas formas. Apesar da complexidade inerente a um modelo de dados tão completo, é possível apresentar apenas uma parte do modelo através da personalização dos formulários de entrada de dados e esconder as porções do modelo que não são utilizadas pela coleção. No outro extremo do espectro, a flexibilidade do Specify permite adicionar campos aos formulários de dados, o que possibilita a sua adequação a diferentes tipos de coleções.
- Interface com alta capacidade de personalização. É possível modificar todos os formulários de entrada de dados do *software* e personalizá-los para as necessidades específicas de cada coleção. O *software* permite a inclusão ou remoção de campos existentes nas várias tabelas do modelo de dados, de forma a otimizar a interface para que apresente apenas os campos relevantes para uma dada coleção. O *software* pode ser pré-configurado para cada uma das disciplinas suportadas (botânica, ictiologia, herpetologia, zoologia em geral, paleontologia), tendo um manual específico para cada uma.
- Documentação rica, completa e didática. Além do manual do *software* existem manuais voltados para cada tipo de coleção em especial.
- É gratuito.
- Suporte técnico gratuito e completo. A equipe do Specify oferece suporte técnico para tarefas como a instalação e configuração do *software*, a personalização dos formulários de entrada de dados, a importação dos dados de outros sistemas ao Specify e resolução de problemas em geral.

Foram identificados como pontos negativos:

- Estabilidade. O sistema deixou de responder ao usuário por algumas vezes durante a avaliação.
- Há algumas pequenas inconsistências na interface, quando por exemplo uma janela ocasionalmente não fecha após a realização de uma tarefa.
- Impossibilidade de importação e exportação dos dados pelos próprios usuários. Devido à complexidade do modelo de dados, é impossível importar dados diretamente no sistema sem auxílio do gerente de dados da equipe de desenvolvimento do Specify.

3.1.4. MICROSOFT EXCEL

Apesar de não ser um *software* específico para o gerenciamento de coleções é muito utilizado para manter os dados de acervos pequenos. As principais razões para tal tendência são a ampla disponibilidade do *software* (parte do pacote de ferramentas Microsoft Office) e a facilidade de aprendizado por ter uma interface intuitiva que facilita a entrada de dados de coleções.

[5] http://usobi.org/specify/

São raras as coleções que não tenham iniciado o processo de informatização, através da iniciativa individual de pesquisadores, preenchendo planilhas MS-Excel com os dados de suas sub coleções.

Apesar de estar amplamente disponível a uma grande parte dos usuários e curadores de coleções e ter uma interface gráfica bastante intuitiva, chegou-se à conclusão que este não é um *software* apropriado para a informatização de coleções. Entre os principais pontos negativos, podem ser citados:

- Não há mecanismos que garantam a integridade dos dados.
- Possui um limite máximo de registros (65 mil).
- Não há suporte ao paradigma relacional, ou seja, não é possível criar relacionamentos entre grupos de dados. No MS-Excel, é impossível expressar de maneira simples o relacionamento de um elemento para muitos, como é o caso de bancos de dados relacionais.
- Não há garantias de que os dados informatizados no MS-Excel possam ser importados sem erros para outros *software* mais apropriados.

3.1.5. MICROSOFT ACCESS E SISTEMAS GERENCIADORES DE BANCOS DE DADOS RELACIONAIS

O MS-Access é um sistema de bancos de dados leve projetado para ser usado por usuários que não sejam exatamente técnicos em informática. Exige um nível de treinamento bem mais elevado que o MS-Excel, e um conhecimento básico do paradigma de bancos de dados relacionais.

Entretanto, à medida que novas características são adicionadas ao banco de dados através de novas tabelas e relacionamentos, cresce também a complexidade no seu desenvolvimento, alimentação e manutenção. Em geral, este é o motivo que leva a maioria dos curadores a abandonar o *software*.

Em lugares onde existe uma equipe de programadores e técnicos em informática, este *software* pode ser útil para a criação de soluções personalizadas para coleções de pequeno porte.

Para coleções de maior porte, é indicado o uso de Sistemas Gerenciadores de Bancos de Dados Relacionais, como o Oracle, Microsoft SQL Server e PostgreSQL, entre outros, assim como uma plataforma de desenvolvimento de aplicativos específica. Nesse caso, é necessária a presença de uma equipe de programadores e analistas de sistema para que se possa desenvolver o sistema apropriado para as necessidades da instituição e de suas coleções.

3.2. INTERAÇÃO COM AS COLEÇÕES

Grande parte do esforço realizado no projeto foi empregado na interação com as coleções científicas participantes. Além da atividade básica, que é a conexão dos bancos de dados à rede do projeto, algumas coleções necessitaram de apoio em seu processo de informatização. A seguir apresentamos um breve relato da interação com as coleções.

3.2.1. COLEÇÃO BRASILEIRA DE MICRORGANISMOS DE AMBIENTE E INDÚSTRIA, CBMAI - CPQBA/UNICAMP

Descrição

A coleção possui linhagens de arqueas, bactérias, fungos filamentosos, plasmídeos e organismos geneticamente modificados (OGM), de interesse industrial e ambiental.

Informatização e conectividade

A coleção está totalmente informatizada utilizando software próprio, especialmente desenvolvido para o gerenciamento da coleção em ambiente Windows/MS-Access. Possui ótima conectividade, fazendo parte da rede da Unicamp.

Integração ao speciesLink

Essa coleção participa do projeto SICol[6] e, por isso, já envia seus dados para o CRIA periodicamente através desse sistema. Assim, toda atualização feita no SICol é refletida em tempo real no speciesLink. Os dados são espelhados no Servidor Local do CRIA. Disponibiliza hoje cerca de 110 registros.

3.2.2. COLEÇÃO DE CULTURAS DE FITOBACTÉRIAS DO LABORATÓRIO DE BACTERIOLOGIA VEGETAL, IBSBF - INSTITUTO BIOLÓGICO DE CAMPINAS

Descrição

Contém cerca de 2.000 linhagens de bactérias fitopatogênicas de interesse para estudos epidemiológicos de fitobacterioses, sendo que 80 são linhagens tipo/patotipo. A maior parte do acervo é composta pelos gêneros *Agrobacterium*, *Clavibacter*, *Curtobacterium*, *Enterobacter*, *Pseudomonas*, *Erwinia*, *Ralstonia*, *Xanthomonas* e *Xylella*.

Informatização e conectividade

A coleção está totalmente informatizada utilizando planilhas MS-Excel. Possui acesso Internet através de linha dedicada a 64Kbps.

Integração ao speciesLink

Esta coleção também participa do projeto SICol enviando seus dados periodicamente para o CRIA através desse sistema. Toda atualização é refletida em tempo real no speciesLink. Os dados são espelhados no Servidor Local do CRIA. Disponibiliza hoje um total de 929 registros.

3.2.3. AS COLEÇÕES DO HERBÁRIO DO ESTADO "MARIA ENEYDA P. KAUFMANN FIDALGO", SP - INSTITUTO DE BOTÂNICA

a. Coleção de Algas

Descrição

A coleção de algas possui cerca de 5.000 espécimens de algas microscópicas e 10.000 espécimens de algas macroscópicas.

[6] http://sicol.cria.org.br/

Informatização e conectividade

Cerca de 7.000 registros de algas estavam armazenados em MicroIsis, um *software* de gerenciamento de bibliotecas e trabalhava em ambiente DOS. Os arquivos estavam armazenados em disquetes 5 1/4 e datavam de 1992. A equipe do CRIA conseguiu uma cópia do MicroIsis na plataforma Windows e recuperou os dados que foram então exportados para MS-Excel.

Os dados do acervo estão em MS-Access, sendo que a instituição está estudando a possibilidade de adotar um *software* mais específico de gerenciamento da coleção. Atendendo ao pedido da curadoria, a equipe do CRIA apresentou o Brahms e detalhou aspectos relevantes do *software*. Foi feita a importação dos dados para o Brahms e instalada uma cópia do *software* no computador da coleção. Por fim o auxiliar de informatização recebeu treinamento no uso do Brahms.

Estão conectados à Internet através da rede da Cetesb, com um link bastante precário e utilizando um servidor Proxy, o que torna a conexão um pouco mais complicada.

Integração ao speciesLink

A coleção estará sendo integrada ao sistema tão logo esteja com o processo de informatização definido e deverá ser espelhada no Servidor Regional de São Paulo.

b. Coleção de Fungos

Descrição

A coleção de fungos possui cerca de 2.500 linhagens, representando cerca de mil espécies.

Informatização e conectividade

Assim como na coleção de algas, está sendo estudada a possibilidade de os dados serem transferidos para o Brahms.

Integração ao speciesLink

A coleção estará sendo integrada ao sistema tão logo esteja com o processo de informatização definido e deverá ser espelhada no Servidor Regional de São Paulo.

c. Coleção de Fanerógamas

Descrição

O Herbário possui mais de 350.000 exsicatas documentando a flora brasileira, das quais 250.000 são fanerógamas. É o melhor documentário da biodiversidade da flora paulista com representantes de todos os grupos vegetais. O Herbário está entre os três maiores do Brasil, em número de exsicatas catalogadas.

Informatização e conectividade

O Herbário encontra-se em processo de informatização utilizando o *software* Brahms.

O processo de informatização se iniciou com o banco de dados particular da pesquisadora Maria Cândida Mamede (Malpighiaceae) e está em andamento. O projeto alocou um bolsista de treinamento técnico nível 3 para auxiliar na informatização e em

breve, a coleção dos materiais “tipo” será disponibilizada no speciesLink. A Família Euphorbiaceae está sendo verificada e informatizada pela Dra. Inês Cordeiro, juntamente com o bolsista do projeto.

A conexão com a Internet é muito lenta e, em horário de pico, praticamente inoperante. Seria necessário um maior investimento na infra-estrutura física do herbário, com uma reformulação da rede interna e da conexão com a Internet. Dada a importância e o tamanho do acervo (350 mil exsicatas), é necessário planejar o processo de informatização da coleção com os recursos adequados (*hardware*, *software* e equipe).

Integração ao speciesLink

A coleção estará sendo integrada ao sistema em breve e deverá ser espelhada no Servidor Regional de São Paulo.

3.2.4. HERBÁRIO DA UNIVERSIDADE ESTADUAL DE CAMPINAS, UEC - IB/UNICAMP

Descrição

O acervo do herbário da Unicamp é o terceiro maior do estado para Fanerógamas. Possui cerca de 130.000 exsicatas, 448 fotografias, 156 materiais-tipo principalmente do Estado de São Paulo. São coletas do cerrado e matas das Regiões Sudeste, Sul e Centro-Oeste, provenientes de projetos de florística e fitossociologia.

Informatização e conectividade

O herbário está em processo de informatização de seu acervo. No início o herbário contava com um computador cedido pelo projeto e dois bolsistas de treinamento técnico nível III. Foram utilizadas planilhas MS-Excel para digitalizar e armazenar as informações. Devido ao acúmulo de dados e à crescente necessidade de padronização da informação e gerenciamento, optou-se pelo *software* Biota. A equipe CRIA importou os dados da planilha MS-Excel para o banco de dados Biota *single-user*, licença que era economicamente viável ao projeto. Porém, o herbário necessitaria da versão multiusuário (que é mais cara) e uma análise do custo-benefício fez com que esse *software* fosse descartado.

A coleção decidiu voltar a utilizar as planilhas MS-Excel enquanto buscava outras ferramentas que melhor se adequassem à sua rotina de trabalho. A equipe do CRIA apresentou o *software* de gerenciamento Brahms para os responsáveis da coleção. Por entender que o *software* atenderia às necessidades do herbário e pelo fato dele ser utilizado por outros herbários no Brasil, o Brahms está sendo adotado.

A equipe do CRIA importou os dados da planilha MS-Excel para o Brahms e ofereceu treinamento intensivo ao bolsista responsável pela administração do banco de dados. Atualmente, todos os dados do herbário estão armazenados no banco Brahms e já estão sendo migrados para o servidor regional do sistema distribuído do speciesLink para busca. Essa migração é de total responsabilidade da coleção.

A equipe do herbário da Unicamp está iniciando a etapa de verificação da qualidade dos dados em paralelo ao processo de informatização. Muitos erros podem ocorrer no processo de digitação e as etiquetas podem conter informações desatualizadas, erradas ou incompletas (taxonomia, localidade, coordenadas, etc.). Ferramentas que possam auxiliar esse processo estão sendo pesquisadas pelo CRIA. Também está sendo implementado o uso de código de barras para a etiqueta na exsicata. Esse mecanismo visa facilitar a rotina de empréstimos de material.

O CRIA cedeu por comodato mais um computador ao herbário da Unicamp (além daquele adquirido pelo projeto) para auxiliar na informatização do acervo. Assim o herbário conta com 4 computadores, uma bolsista de pós-doutorado (qualidade dos dados), dois bolsistas de treinamento técnico nível III (administração do banco e digitação) e três estudantes com bolsa de trabalho da Unicamp (digitação).

Integração ao speciesLink

Os dados são espelhados no Servidor Regional de Campinas. Disponibiliza hoje 12.860 registros.

3.2.5. HERBÁRIO DO DEPARTAMENTO DE BOTÂNICA, SPF - IB/USP

Descrição

O herbário possui cerca de 133.500 exsicatas, 460 amostras na carpoteca, 1.200 amostras na xiloteca, 420 espécimens em meio líquido, 325 fotografias e 460 materiais-tipo. O acervo contém as seguintes coleções especiais: Flora da Serra do Cipó (MG), Flora de Grão Mogol (MG), Flora de Campos Rupestres (MG, BA), Flora de Catolés e Pico das Almas (BA); e coleções de Wilson Hoehne e Aylthon B. Joly.

Informatização e conectividade

A coleção adotou o *software* Brahms para gerenciar seus dados. O único conjunto de dados informatizados da coleção é o de algas, contendo cerca de 22.000 registros de todo o Brasil. Esses registros estavam em planilhas MS-Excel e foram importados para o banco de dados do Brahms. O *software* foi instalado na máquina da coleção e os responsáveis pela informatização foram treinados para inserir e gerenciar os dados.

Para acelerar o processo seria necessário que essa coleção tivesse um maior investimento na infra-estrutura computacional e em mão-de-obra para a digitação de seu acervo.

Integração ao speciesLink

Os dados da sub-coleção de algas estão disponíveis no sistema através do Servidor Regional de Campinas. Tão logo o Servidor Regional de São Paulo seja instalado, este será o espelho da coleção. São 20.897 registros *on-line*.

3.2.6. COLEÇÃO DE ÁCAROS DO DEPARTAMENTO DE ENTOMOLOGIA, FITOPATOLOGIA E ZOOLOGIA, ACARIESALQ - LEF/ESALQ

Descrição

A coleção de ácaros da ESALQ possui cerca de 15.000 exemplares catalogados, dos quais cerca de 1.000 correspondem a tipos de 200 espécies. A coleção tem ácaros de interesse agrícola, constituídos por fitófagos e predadores. A coleção é composta em sua maioria de ácaros do Estado de São Paulo, embora contenha também uma grande quantidade de exemplares de outros estados brasileiros e de países de todos os continentes. A coleção está utilizando o *software* Biota para gerenciar os dados do acervo.

Informatização e conectividade

A Coleção de Ácaros da ESALQ é a única coleção contatada que possui um Gerente de Coleção, isto é, uma pessoa contratada exclusivamente para cuidar da coleção. Isso facilitou muito o trabalho da equipe CRIA por ter um contato permanente na coleção. O Gerente de Coleção recebeu treinamento para inserção de dados no Biota e no uso do *spLinker,* software que realiza a migração dos dados do Biota para o Servidor Regional do *species*Link.

A coleção tem um sistema de captura de imagens digital de microscópio (vídeo) e gostaria de enviar as imagens pela Internet para identificação de ácaros à distância. Especialistas poderiam ver as imagens em tempo real e solicitar foco e ângulo diferentes.

Também está sendo realizado com pesquisadores da coleção um projeto de modelagem do ácaro vermelho do tomate. Essa é uma praga agrícola que proporciona grandes perdas da produção não só no Brasil, mas no mundo. O objetivo do trabalho é dimensionar o nicho potencial do ácaro e determinar áreas onde existe maior probabilidade de ocorrência de inimigos naturais para estudos de controle biológico.

Em uma próxima etapa do projeto, seria oportuno o estudo de transmissão de imagens *on-line* pela Internet e a reestruturação do banco Biota para permitir a vinculação de hospedeiros aos ácaros.

Integração ao speciesLink

A coleção está disponível no sistema do Servidor Regional de Campinas. Disponibiliza hoje 12.392 registros.

3.2.7. COLEÇÃO DE ÁCAROS DO DEPARTAMENTO DE ZOOLOGIA E BOTÂNICA, ACARIDZSJRP - IBILCE/UNESP

Descrição

A coleção possui cerca de 7.000 exemplares, principalmente do noroeste do Estado de São Paulo, mas possui também amostras de exemplares da Argentina, Colômbia e Indonésia, e parátipos doados (8) de 2 espécies africanas (6) e Filipinas (2).

Informatização e conectividade

A coleção utilizava o *software* MS-Excel para gerenciar a coleção, mas, por sugestão do CRIA, o acervo foi migrado para o *software* de gerenciamento Biota. Foi realizada a importação dos dados do MS-Excel para o Biota e foi dado um treinamento no CRIA ao bolsista da coleção. Além de acompanhar a formatação, estruturação e importação dos dados, o bolsista ficou responsável por disseminar o aprendizado para os pesquisadores e outros usuários da coleção. Os dados foram estruturados para facilitar buscas tanto de ácaros quanto dos hospedeiros o que facilita os estudos de controle de pragas.

Devido à distância física entre a coleção e o CRIA, para facilitar a comunicação foi instalado o NetMeeting e uma câmera digital no computador da coleção. Esse equipamento permite a realização de vídeo conferências o que foi possibilitou oferecer suporte à distância.

Integração ao speciesLink

A coleção está disponível através do Servidor Regional de São José do Rio Preto instalado no Pólo Computacional do campus da Unesp com conexão com a Internet 2. Disponibiliza hoje 5.265 registros.

3.2.8. Coleção de Peixes do Departamento de Zoologia e Botânica, DZSJRP - IBILCE/UNESP

Descrição

A coleção possui 5.000 lotes, com aproximadamente 23.000 exemplares. Os exemplares são provenientes em sua grande maioria da região Noroeste do Estado de São Paulo, drenagens dos rios Turvo/Grande e médio/baixo Tietê, e representam uma excelente cobertura para essas áreas. Abriga ainda amostras dos rios litorâneos da região Sudeste brasileira (principalmente Ribeira de Iguape), do rio São Francisco e rios amazônicos.

Informatização e conectividade

O gerenciamento da coleção já era feito utilizando o *software* Biota, sendo que 5.483 registros estão disponíveis no sistema.

Foi instalado o sistema de vídeo conferência via webcam e controle remoto de desktop via NetMeeting.

A coleção está muito bem estruturada. O curador, Dr. Francisco Langeani, tem uma excelente preocupação com a qualidade dos dados e procura sempre estar atualizado com novas ferramentas de informática.

Integração ao speciesLink

A coleção está disponível através do Servidor Regional de São José do Rio Preto. Disponibiliza hoje 5.491 registros.

3.2.9. Coleção de Peixes do Laboratório de Ictiologia de Ribeirão Preto, LIRP - FFCLRP/USP

Descrição

A coleção contém cerca de 30.000 exemplares, principalmente de água doce. Os exemplares são provenientes de locais muito pouco coletados da Floresta Costeira Atlântica do leste do Brasil; cabeceiras do rio São Francisco, em Minas Gerais; bacia do rio Pardo, nas imediações de Ribeirão Preto, SP; da ASPE-SEMA do CEBIMar-USP, localizada à margem continental do Canal de São Sebastião, SP; e de riachos das bacias do São José dos Dourados, Baixo Tietê, Aguapeí e Peixe, todas do sistema do Alto rio Paraná, no Estado de São Paulo. A coleção não contém tipos.

Informatização e conectividade

Possui cerca de 4.500 registros, todos informatizados utilizando o *software* Biota. A coleção já possuía um computador apenas para a informatização e gerenciamento dos dados com o Biota. Optaram por dedicar a máquina do projeto para ser um espelho do banco e estar conectada diretamente na rede, tendo a configuração de um Servidor Local (do LIRP).

Integração ao speciesLink

Está disponível através do Servidor Local de LIRP. Disponibiliza hoje 4.314 registros.

3.2.10. COLEÇÃO DE PEIXES DO MUSEU DE ZOOLOGIA, MZUSP - IB/USP

Descrição

A coleção possui aproximadamente 60.000 lotes e cerca de 700.000 exemplares catalogados e cerca de 24.000 lotes e 200.000 exemplares não catalogados. É a maior coleção de peixes amazônicos do mundo. Possui material tipo de aproximadamente 400 táxons, compreendendo cerca de 1.000 lotes.

Informatização e conectividade

Está 100% informatizada, usando o *software* MUSE e migrando para o *software* TEIA, a ser implantado em todo o Museu.

O TEIA é um *software* que foi produzido por uma empresa contratada para gerenciar os dados de todas as coleções do Museu. Até a última visita técnica ele ainda não tinha sido implementado. O coordenador técnico da coleção Osvaldo Takeshi Oyakawa manifestou o interesse em migrar os dados da coleção do MUSE para o Specify após a demonstração do *software* pela equipe CRIA. A equipe CRIA fez o primeiro contato com a equipe da Universidade de Kansas, responsável pelo desenvolvimento do *software* (Specify), para que ela desse suporte ao museu na importação dos dados. Fomos informados que os dados foram enviados e o processo de importação no Specify está sendo desenvolvido.

Integração ao speciesLink

Será integrada ao sistema através do Servidor Regional de São Paulo tão logo o processo de informatização dos dados esteja definido.

Além das coleções previamente selecionadas quando da concepção do projeto, estão também em fase de organização de seus dados para ingresso no speciesLink os herbários do Instituto Agronômico de Campinas, IAC, e da ESALQ. Foram também feitos os primeiros contatos com os herbários da Universidade Estadual Paulista “Julio de Mesquita Filho” UNESP – campus de Botucatu (BOTU) e do Instituto Florestal de São Paulo (SPSF).

3.3. ARQUITETURA DA REDE DO SPECIESLINK

Todo o *software* desenvolvido para o speciesLink funciona sobre o sistema operacional Linux usando apenas ferramentas de *software* livre. Podemos definir de forma simplificada, *software* livre como o *software* cujo autor o distribui e outorga a todos a liberdade de uso, cópia, alteração e redistribuição de sua obra. A liberdade de uso e alteração é viabilizada através da distribuição dos códigos fonte do programa. Além do código fonte, o autor do programa outorga a liberdade para que outros programadores possam modificar o código original e redistribuir versões modificadas. Dentre as vantagens decorrentes da utilização de *software* livre destacamos: custo social baixo, pois não se fica refém da tecnologia proprietária; independência de fornecedor único; desembolso inicial próximo de zero e a não obsolescência do *hardware*. Estas características do *software* livre são extremamente favoráveis ao desenvolvimento de

sistemas de informação voltados à comunidade científica, uma vez que o parque de máquinas instalado não necessita ser atualizado com a freqüência que seria necessária no caso da utilização de *software* proprietário, visto que este induz à aquisição continuada de novas plataformas.

Em particular, os bancos de dados são implementados no sistema gerenciador de bancos de dados PostgreSQL[7] e o *software* foi desenvolvido usando-se as linguagens PHP[8], Perl[9] e Java[10]. O sistema também faz uso dos *software* Apache (como servidor de páginas web) e Tomcat[11] (como servidor de páginas dinâmicas para o portal). A experiência que o CRIA já possuía no desenvolvimento de *software* usando ferramentas de código livre (como por exemplo, o *Sin*Biota) tornou bastante confortável essa opção com a certeza que as ferramentas teriam desempenho compatível com as necessidades do projeto e muitas vezes superior ao desempenho que seria obtido com a utilização de *software* proprietário. A opção por *software* livre também foi feita pensando-se na ampliação da rede a um custo baixo, sem demandar a compra de licenças.

A plataforma de *hardware* escolhida (Linux + Intel) deve-se à grande confiabilidade que a mesma tem demonstrado nos últimos anos, aliado ao custo baixo e ao desempenho compatível com plataformas proprietárias mais caras (por exemplo, a plataforma Sun + Solaris).

A rede de coleções do projeto speciesLink utiliza o protocolo DiGIR para integrar e oferecer a possibilidade de rastreamento de informações em bancos de dados distintos, apresentando os resultados ao usuário como se as informações tivessem origem numa base de dados única.

Foi criado um esquema conceitual para que a rede atendesse às necessidades da comunidade local e ao mesmo tempo fosse compatível com os esquemas usados pelas outras redes internacionais para permitir a integração com estas redes.

O esquema conceitual usado é derivado do esquema Darwin Core 2.0, disponível no website projeto Species Analyst[12].

[7] http://www.postgresql.org
[8] http://www.php.net
[9] http://www.perl.org
[10] http://java.sun.com
[11] http://www.apache.org
[12] Fonte: http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV2

A implantada tem tem a seguinte topologia.

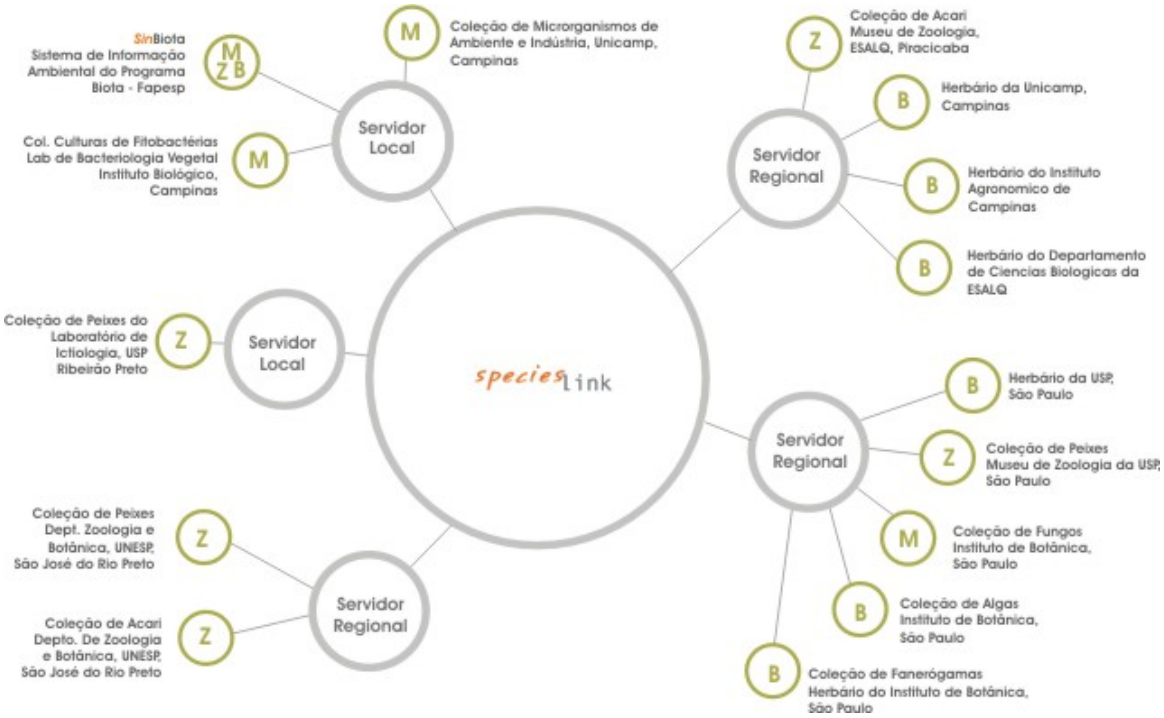


Figura 1. Esquema da rede speciesLink

Uma descrição detalhada do protocolo utilizado e a definição dos "servidores" apresentados no diagrama serão apresentadas a seguir.

3.4. DESENVOLVIMENTO DE UM PROTOCOLO CLIENTE/SERVIDOR PARA RECUPERAR INFORMAÇÃO DE FONTES DISTRIBUÍDAS

No início do desenvolvimento do trabalho foi lançado um projeto cooperativo *on-line* denominado DiGIR, Distributed Generic Information Retrieval[13]. A equipe resolveu participar do desenvolvimento desse projeto ao invés de criar um sistema próprio por entender que havia a possibilidade de ser criada uma ferramenta que poderia vir a se tornar um padrão e também pela possibilidade de se beneficiar do conhecimento de desenvolvedores de outros países, ampliando, dessa forma a equipe de desenvolvimento. A experiência tem sido muito bem sucedida.

O DiGIR é um protocolo cliente/servidor projetado para recuperar informação de fontes distribuídas. Usa o protocolo Hypertext Transport Protocol (HTTP) como mecanismo de transferência de dados e o Extensible Markup Language (XML) para codificar as mensagens trocadas entre clientes e servidores. Foi projetado para suportar a recuperação de informação conforme um modelo de dados genérico e arbitrário. O protocolo mantém a independência entre o mecanismo de transmissão de mensagens e o modelo de dados em que a informação é recuperada. Dessa forma é possível utilizar

[13] http://sourceforge.net/projects/digir/

o protocolo para recuperar dados de outros domínios e não apenas de coleções biológicas.

A maneira mais comum de utilizar o protocolo é feita através de três componentes principais: uma camada de apresentação (responsável pela interação com o usuário final), um portal (responsável pela distribuição das consultas), e os provedores de dados (responsáveis por interagir com cada banco de dados ligado à rede). A interface entre os três componentes é muito bem definida, permitindo seu desacoplamento e também possibilitando implementações em diferentes linguagens de programação e plataformas computacionais.

Entre as principais características do protocolo encontram-se:

- **Transparência de localização:** a complexidade de identificação da localização dos vários provedores de dados disponíveis na rede função do portal;
- **Transparência de plataforma e modelo local de dados:** cada provedor de dados é responsável pela tradução necessária entre o seu próprio modelo de dados (esquema do banco de dados) e o modelo genérico adotado pela comunidade a que este pertence;
- **Descoberta automática de fontes de dados:** através do Universal Description, Discovery and Integration (UDDI), o portal DiGIR pode detectar automaticamente quais são os provedores servindo dados segundo um determinado modelo. Os provedores, por sua vez, ao serem configurados, podem registrar-se em um dos diretórios UDDI públicos disponíveis para que os portais existentes possam descobri-los e associá-los às suas redes.

3.4.1. PROTOCOLO DE COMUNICAÇÃO

O protocolo DiGIR é especificado através de um esquema XML (*XML Schema*) que define a estrutura das mensagens enviadas e retornadas por provedores DiGIR[14].

Para desacoplar o protocolo DiGIR do modelo de dados usado por uma rede DiGIR, são definidos elementos de busca genéricos e fictícios que são substituídos em um outro esquema XML que define o modelo de dados a ser compartilhado por cada comunidade em particular[15]. Para isso são usados dois esquemas XML separados e hierárquicos, um para o protocolo, definindo a estrutura das mensagens trocadas entre os componentes da rede e outro para a definição dos conceitos específicos para cada domínio de informação. Nesta seção é descrita a estrutura do primeiro esquema, o esquema XML do protocolo DiGIR.

O esquema XML do protocolo define duas estruturas principais: o elemento <*request*> que contém uma consulta a um provedor DiGIR e o elemento <*response*> que define a estrutura da resposta do provedor à consulta.

O elemento <*request*> contém um elemento <*header*> onde são definidos: a versão do protocolo, o horário em que a consulta está sendo feita, o endereço IP de quem originou a consulta (no caso o IP registrado é o do usuário final para efeito de registro de atividade do sistema), o(s) destinatário(s) da mensagem (podem ser um único provedor ou múltiplos) e o tipo de consulta que está sendo realizada.

[14] http://digir.sourceforge.net/prot/beta3/digir.xsd
[15] http://digir.sourceforge.net/fed/beta3/darwin2.xsd

Atualmente o protocolo DiGIR suporta três tipos de consulta: pedido de metadados, consulta a registros de dados e pedido de inventário.

Metadados

O pedido de metadados é especificado quando o elemento <*type*> contém o valor "*metadata*", não sendo necessária a existência do corpo da mensagem.

Abaixo é apresentado um exemplo de pedido de metadados no DiGIR:

```
<request xmlns="http://www.namespaceTBD.org/digir">
 <header>
 <version>>$Revision: 1.7 $</version>
 <sendTime>2003-06-04 18:07:34-0300</sendTime>
 <source>200.144.120.37</source>
 <destination>http://splink.cria.org.br/provider/DiGIR.php
 </destination>
 <type>metadata</type>
 </header>
</request>
```

Este pedido obtém como resposta um documento XML descrevendo cada provedor especificado como destinatário da mensagem e cada coleção (*resource*) conectada ao provedor:

```
<response xmlns="http://digir.net/schema/protocol/2003/1.0">
 <header>
 <version>$Revision: 1.7 $</version>
 <sendTime>2003-06-04 18:07:35-0300</sendTime>
 <source>http://splink.cria.org.br:80/provider/DiGIR.php</source>
 <destination>200.144.120.37</destination>
 </header>
 <content>
 <metadata>
  <provider>

  <name>speciesLink_CRIA</name>

<accessPoint>http://splink.cria.org.br:80/provider/DiGIR.php</acc
essPoint>
  <implementation>$Revision: 1.7 $</implementation>
  <host>
   <name>Centro de Referência em Informação Ambiental</name>
   <code>CRIA</code>

<relatedInformation>http://www.cria.org.br/</relatedInformation>
   <contact type="technical">
   <name>Sidnei de Souza</name>
   <title>Analista de Sistemas</title>
   <emailAddress>sidnei@cria.org.br</emailAddress>
   <phone>+55 19 32880466</phone>
   </contact>
   <abstract>O Centro de Referência em Informação Ambiental, é
uma sociedade civil, sem fins lucrativos, que pretende disseminar
o conhecimento científico e tecnológico e promover a educação,
visando a conservação e a utilização sustentável dos recursos
naturais e a formação da cidadania.</abstract>
  </host>
  <resource>
```

```
   <name>Coleção Brasileira de Microrganismos de Ambiente e
Indústria</name>
   <code>CBMAI</code>

<relatedInformation>http://www.cpqba.unicamp.br/</relatedInformat
ion>
   <contact type="administrative ">
   <name>Gilson Paulo Manfio</name>
   <title>Curador</title>
   <emailAddress>gmanfio@cpqba.unicamp.br</emailAddress>
   <phone>+55 19 38847500</phone>
   </contact>
   <abstract>A coleção possui cerca de 700 linhagens de arqueas,
bactérias, fungos filamentosos, plasmídeos e organismos
geneticamente modificados (OGM), de interesse industrial e
ambiental. As principais atividades em que a coleção é utilizada
são: identificação, treinamento, pesquisa e acessoria à
indústria.</abstract>
   <keywords>bactérias, fungos</keywords>
   <citation />
   <useRestrictions />
   <conceptualSchema
schemaLocation="http://digir.sourceforge.net/fed/beta3/darwin2.xs
d">http://www.namespaceTBD.org/darwin2</conceptualSchema>
   <recordIdentifier>CBMAI</recordIdentifier>
   <recordBasis>strain</recordBasis>
   <numberOfRecords>110</numberOfRecords>
   <dateLastUpdated>10/04/2003</dateLastUpdated>
   <minQueryTermLength>0</minQueryTermLength>
   <maxSearchResponseRecords>1000</maxSearchResponseRecords>

<maxInventoryResponseRecords>1000</maxInventoryResponseRecords>


      <defaultRecordFormat>http://digir.sourceforge.net/prov/darw
      in/darwin2brief.xsd</defaultRecordFormat>

<defaultInventoryConcept>darwin:ScientificName</defaultInventoryC
oncept>
  </resource>
  </provider>
 </metadata>
 </content>
 <diagnostics>
 <diagnostic code="STATUS_INTERVAL"
severity="info">3600</diagnostic>
 <diagnostic code="STATUS_DATA"
severity="info">1,3,0</diagnostic>
 </diagnostics>
</response>
```

Registros

Uma consulta a registros de dados é especificada pelo elemento <*search*> dentro da requisição <*request*>. O elemento <*search*> contém dois outros elementos: um elemento <*filter*> e outro <*records*>. O primeiro define a consulta a ser realizada e o segundo define quais campos devem ser retornados na resposta. Numa comparação

com uma consulta SQL, o elemento <*filter*> seria equivalente à cláusula *WHERE* e o <*records*> à projeção ou cláusula *SELECT*.

O filtro (estrutura definida no elemento <*filter*>) define uma consulta a ser realizada no provedor através da combinação de operadores lógicos (LOPs) e comparativos (COPs) numa estrutura em árvore expressa em XML. Os operadores de comparação são usados para comparar o valor de um campo com uma constante enquanto os operadores lógicos possibilitam a comparação lógica entre dois operadores de comparação ou dois operadores lógicos.

Os operadores lógicos utilizados são: *E lógico*, *OU lógico* e a *negação*.

Os operadores comparativos utilizados são: *menor que* (<), *menor ou igual a* (≤), *igual a* (=), *maior que* (>), *maior ou igual* a (≥), *diferente* de (≠), *contém (in), e assemelha-se a (like)*.

O filtro tem a vantagem de ser uma estrutura independente de linguagem que pode ser facilmente transformada em uma consulta SQL e pode ser analisada sintaticamente através de um dos muitos *parsers* XML disponíveis.

O elemento <*records*> é utilizado para definir a estrutura e os tipos de dados a serem retornados na consulta. Este elemento contém uma lista de elementos presentes no esquema XML conceitual, sendo que cada um define um campo a ser retornado na resposta à consulta. O elemento <*records*> pode opcionalmente especificar uma *URL* que contém o documento que descreve o conjunto de registros a serem retornados na consulta.

Abaixo é mostrado um exemplo de consulta do tipo *search*:

```
<request xmlns:digir="http://www.namespaceTBD.org/digir"
 xmlns:darwin="http://www.namespaceTBD.org/darwin2"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema"
 xmlns="http://www.namespaceTBD.org/digir">
<header>
 <version>>$Revision: 1.7 $</version>
 <sendTime>2003-06-05 11:40:50-0300</sendTime>
 <source>129.237.201.120</source>
 <destination
resource="IBSBF">http://splink.cria.org.br/provider/DiGIR.php</de
stination>
 <type>search</type>
</header>
<search>
 <filter>
 <like>
  <darwin:Genus>agrobacterium</darwin:Genus>
 </like>
 </filter>
 <records start="0" limit="10" count="0">
 <structure>
  <xsd:element name="record">
  <xsd:complexType>
   <xsd:sequence>
   <xsd:element ref="darwin:InstitutionCode"/>
   <xsd:element ref="darwin:CollectionCode"/>
   <xsd:element ref="darwin:CatalogNumber"/>
   <xsd:element ref="darwin:Genus"/>
```

```
    <xsd:element ref="darwin:Species"/>
    </xsd:sequence>
   </xsd:complexType>
   </xsd:element>
  </structure>
  </records>
 </search>
 </request>
```

Outras tecnologias foram consideradas para substituir o filtro como linguagem para expressar as consultas. Entre elas foram avaliadas o XPath, o XQuery e o próprio SQL.

O XQuery não era uma tecnologia madura o suficiente para ser utilizada na época da definição do protocolo. Além disso, tanto o XPath quanto o XQuery exigem transformações não triviais para conversão em SQL, linguagem de consulta nativa utilizada pela maioria dos bancos de dados existentes nos provedores. Essas tecnologias seriam mais apropriadas caso os provedores utilizassem bancos de dados XML nativos.

A linguagem SQL foi descartada pois existem muitos dialetos diferentes no mercado, o que exigiria o desenvolvimento de um analisador sintático para extrair os componentes da consulta e reconstruir a expressão de busca no dialeto padrão ANSI.

Como exemplo de resposta a uma consulta de registros, temos:

```
<response xmlns="http://digir.net/schema/protocol/2003/1.0">
 <header>
 <version>$Revision: 1.7 $</version>
 <sendTime>2003-06-05 11:40:51-0300</sendTime>
 <source
resource="IBSBF">http://splink.cria.org.br:80/provider/DiGIR.php<
/source>
 <destination>129.237.201.120</destination>
 </header>
 <content xmlns:darwin="http://www.namespaceTBD.org/darwin2"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
 <record>
  <darwin:InstitutionCode>IB</darwin:InstitutionCode>
  <darwin:CollectionCode>IBSBF</darwin:CollectionCode>
  <darwin:CatalogNumber>343</darwin:CatalogNumber>
  <darwin:Genus>Agrobacterium</darwin:Genus>
  <darwin:Species>radiobacter</darwin:Species>
 </record>
 <record>
  <darwin:InstitutionCode>IB</darwin:InstitutionCode>
  <darwin:CollectionCode>IBSBF</darwin:CollectionCode>
  <darwin:CatalogNumber>307</darwin:CatalogNumber>
  <darwin:Genus>Agrobacterium</darwin:Genus>
  <darwin:Species>rhizogenes</darwin:Species>
 </record>
 </content>
 <diagnostics>
 <diagnostic code="STATUS_INTERVAL"
severity="info">3600</diagnostic>
 <diagnostic code="STATUS_DATA"
severity="info">3,13,0</diagnostic>
 <diagnostic code="MATCH_COUNT" severity="info">2</diagnostic>
```

```
 <diagnostic code="RECORD_COUNT" severity="info">2</diagnostic>
 <diagnostic code="END_OF_RECORDS"
severity="info">false</diagnostic>
 </diagnostics>
</response>
```

Inventário

Finalmente, o terceiro tipo de consulta é o inventário, e é definido através do elemento <*inventory*>. Através dessa consulta é possível obter uma lista dos valores únicos presentes em um campo num provedor, e o número de registros de cada valor retornado é equivalente a um comando SQL: *SELECT DISTINCT campo, COUNT(campo) AS total*.

O campo sobre o qual o inventário será feito é definido dentro do elemento <*inventory*>. Este tem estrutura semelhante ao mesmo elemento no caso da consulta do tipo *search*, mas no caso do inventário, deve conter apenas um registro.

Opcionalmente a requisição de inventário pode conter um filtro, o que permite limitar o alcance do comando a um subconjunto dos registros do provedor.

Abaixo é mostrado um exemplo de consulta do tipo *inventory:*

```
<request xmlns="http://www.namespaceTBD.org/digir"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema"
   xmlns:darwin="http://www.namespaceTBD.org/darwin2">
 <header>
 <version>>$Revision: 1.7 $ </version>
 <sendTime>2003-06-05 11:54:00-03:00</sendTime>
 <source>216.91.87.102</source>
 <destination
resource="IBSBF">http://splink.cria.org.br/provider/DiGIR.php</de
stination>
 <type>inventory</type>
 </header>
 <inventory>
 <darwin:Genus />
 <count>true</count>
 </inventory>
</request>
```

E a seguir uma possível resposta a esta consulta:

```
<response xmlns="http://digir.net/schema/protocol/2003/1.0">
 <header>
 <version>$Revision: 1.7 $</version>
 <sendTime>05-06-2003 11:55:56-0300</sendTime>
 <source
resource="IBSBF">http://splink.cria.org.br:80/provider/DiGIR.php<
/source>
 <destination>216.91.87.102</destination>
 </header>
 <content xmlns:darwin="http://www.namespaceTBD.org/darwin2"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
 <record>
  <darwin:Genus count="33">Agrobacterium</darwin:Genus>
 </record>
 <record>
```

```
  <darwin:Genus count="25">Bacillus</darwin:Genus>
 </record>
 <record>
  <darwin:Genus count="34">Clavibacter</darwin:Genus>
 </record>
 <record>
  <darwin:Genus count="364">Xanthomonas</darwin:Genus>
 </record>
 </content>
 <diagnostics>
 <diagnostic code="STATUS_INTERVAL"
severity="info">3600</diagnostic>
 <diagnostic code="STATUS_DATA"
severity="info">3,14,1</diagnostic>
 <diagnostic code="MATCH_COUNT" severity="info">4</diagnostic>
 <diagnostic code="RECORD_COUNT" severity="info">4</diagnostic>
 <diagnostic code="END_OF_RECORDS"
severity="info">true</diagnostic>
 </diagnostics>
</response>
```

3.4.2. PROVEDOR DE DADOS

Trata-se da camada responsável por abstrair a heterogeneidade das fontes de dados conectadas à rede. Cada provedor de dados pode estar ligado a uma ou mais bases de dados (denominadas "*resources*" pelo protocolo). A comunicação é feita através do protocolo DiGIR. Ao receber uma requisição, o provedor faz a tradução da mesma para o padrão SQL, interage com a base de dados especificada na requisição, e produz uma resposta de acordo com o protocolo DiGIR. Normalmente são portais que se comunicam com provedores de dados.

O *software* utilizado pelo speciesLink para cumprir o papel de provedor de dados faz parte da implementação padrão do protocolo DiGIR, resultado de um esforço internacional já mencionado e do qual nossos desenvolvedores participam ativamente. Seu desenvolvimento foi feito na linguagem PHP[16], por ser multi-plataforma, bem documentada, e particularmente voltada para operar sob HTTP. Além disso, é uma ferramenta de fácil acesso (não proprietária), *open source*, e possui bibliotecas e recursos necessários para trabalhar com XML e abstrair a comunicação com diversos tipos de bases de dados.

Os pré-requisitos para instalação de um provedor de dados são:

- Servidor Web que suporte PHP (já foram testados: Apache e IIS)
- PHP versão maior ou igual a 4.2.3
- Existência de um domínio fixo para o servidor (sem um endereço fixo torna-se impraticável utilizar os serviços de um provedor de dados)
- A configuração do provedor é feita através de arquivos no formato XML. O primeiro deles, normalmente chamado *providerMeta.xml* contém os metadados do provedor. O segundo, chamado *resources.xml*, lista todas as coleções conectadas a ele, indicando para cada uma delas o nome do arquivo de configuração. Este último, além de conter os metadados da coleção, possui

[16] http://www.php.net/

também instruções para conexão com a base de dados e um mapeamento dos campos em relação ao esquema de dados sendo usado pelo protocolo.

3.4.3. MIRROR

Apesar do protocolo DiGIR oferecer bastante flexibilidade em termos de configuração e conexão de bancos de dados de coleções biológicas em diferentes plataformas computacionais, é necessário também que o *hardware* utilizado pelo provedor de dados satisfaça as seguintes condições: i) tenha desempenho suficiente para suportar a carga imposta pelos usuários da rede, e ii) esteja conectado ao portal através de uma linha suficientemente rápida e estável.

Embora várias das coleções biológicas do estado de São Paulo tenham modernizado a sua infra-estrutura computacional e de conectividade em virtude dos investimentos governamentais, principalmente da FAPESP, algumas delas ainda têm poucos recursos. Para mais detalhes sobre a situação específica de cada coleção, veja a seção 3.2.

Visando estender a funcionalidade do protocolo DiGIR para atender a esta situação foi criado pela equipe do CRIA o conceito dos **nós intermediários de espelhamento**. Estes nós intermediários ficam situados entre o provedor de dados DiGIR e o banco de dados da coleção, e servem como espelho dos dados, contendo cópias dos dados da coleção atualizadas periodicamente.

Os nós intermediários funcionam como provedores DiGIR comuns, os quais chamamos de Servidores Regionais ou Locais. Entretanto, ao invés de estarem ligados ao banco de dados de uma coleção em particular, estes estão ligados a um banco de dados especial que armazena espelhos dos dados de uma ou mais coleções. Estes espelhos são atualizados através de um sistema cliente/servidor instalado no nó intermediário (servidor) e no microcomputador da coleção (cliente). A interface de comunicação com o servidor é definida pelos seguintes métodos a serem invocados através do protocolo de comunicação SOAP[17] via HTTP:

GetCollectionId: Retorna a chave primária de uma determinada coleção.

Parâmetros: código da instituição (caracter), código da coleção (caracter).

Valor de retorno: chave primária da coleção (número inteiro maior que zero) ou código de erro.

Reset: Remove todos os registros de uma determinada coleção.

Parâmetros: chave primária da coleção (número inteiro).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

Suspend: Suspende as buscas em dados de uma determinada coleção (função disponível para os administradores da coleção).

Parâmetros: chave primária da coleção (número inteiro).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

[17] http://www.w3.org/TR/SOAP/

Resume: Volta a dispobibilizar os dados de uma determinada coleção que estavam previamente suspensos.

Parâmetros: chave primária da coleção (número inteiro).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

Block: Bloqueia alterações nos dados de uma determinada coleção (funcionalidade disponível para os administradores da rede em caso de manutenção).

Parâmetros: chave primária da coleção (número inteiro).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

Unblock: Desbloqueia dados de uma determinada coleção que estavam previamente bloqueados (funcionalidade disponível para os administradores da rede).

Parâmetros: chave primária da coleção (número inteiro).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

RemoveRecords: Remove registros específicos de uma determinada coleção.

Parâmetros: chave primária da coleção (número inteiro), número total de registros a serem removidos (número inteiro), códigos de identificação dos registros a serem removidos separados por "|" (caracter).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

InsertRecords: Inclui registros de uma determinada coleção.

Parâmetros: chave primária da coleção (número inteiro), número total de registros a serem inseridos (número inteiro), registros a serem inseridos, separados por quebra de linha, com valores separados por "|" (caracter).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

Observação: Caso algum campo contenha o caracter separador ("|"), este deverá vir precedido de barra invertida ("\").

GetCollectionData: Retorna os metadados da coleção que estão armazenados no servidor.

Parâmetros: chave primária da coleção (número inteiro).

Valor de retorno: Metadados da coleção separados por "|" (caracter: código da instituição | nome da instituição | código da coleção | nome da coleção | nome da pessoa para contato | e-mail da pessoa para contato | endereço (url) do site da coleção | código do status da coleção). Ou código de erro.

Observação: Caso algum campo contenha o caracter separador ("|"), este deverá virá precedido de barra invertida ("\").

SetCollectionData: Atualiza os metadados de uma determinada coleção.

Parâmetros: chave primária da coleção (número inteiro), metadados da coleção separados por "|" (código da instituição | nome da instituição | código da coleção | nome da coleção | nome da pessoa para contato | e-mail da pessoa para contato | endereço (url) do site da coleção | código do status da coleção).

Valor de retorno: "1" para indicar sucesso na operação, ou código de erro.

Observações: Caso algum campo contenha o caracter separador ("|"), este deverá vir precedido de barra invertida ("\"). Decidiu-se por descartar os três primeiros parâmetros (só poderão ser alterados pelos administradores da rede).

Códigos de erro:

-1 -> falha na comunicação com o banco de dados

-2 -> operação não realizada

-3 -> parâmetros incorretos

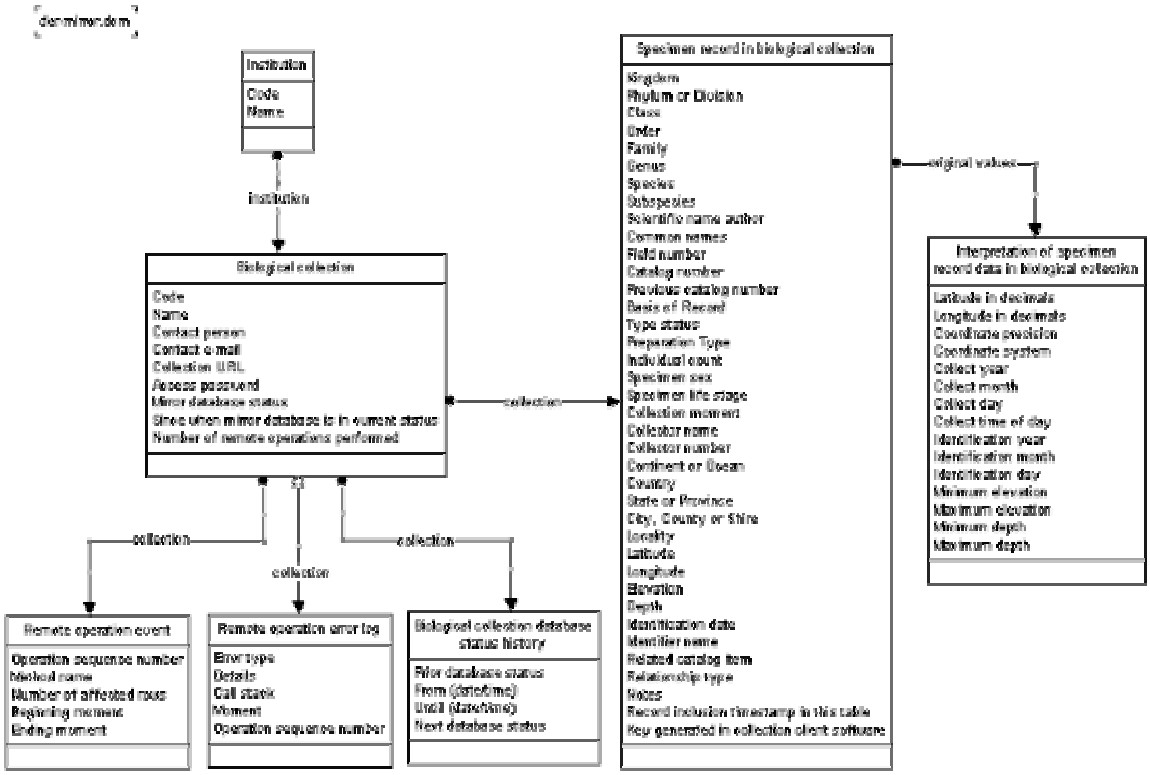
Mais detalhes sobre eventuais erros são armazenados no arquivo de log do servidor.

Nome do domínio para todos os métodos (a ser usado na URI): interface

O servidor foi desenvolvido em linguagem "Perl", por possuir um dos módulos mais bem conceituados para utilização do protocolo SOAP (chamado SOAP::Lite), por ser uma linguagem fortemente especializada em manipulação de texto, e também por ser *open source* e muito bem documentada. O banco de dados escolhido foi o PostgreSQL[18] devido a sua robustez, performance, e disponibilidade de recursos como: controle de transações, manutenção da integridade referencial, e disparadores automáticos.

O diagrama entidade-relacionamento do banco de dados instalado no servidor segue o modelo abaixo:

[18] http://www.postgresql.org/



Ao espelhar os dados de uma coleção, todos os valores originais são armazenados intactos. Entretanto, para alguns campos, faz-se necessária uma interpretação dos dados, pois o protocolo DiGIR também estabelece padrões para alguns deles (ex: latitude e longitude devem estar em decimais, altitude e profundidade em metros, datas são sempre tratadas separadamente em dia, mês e ano). Desta forma, foi incorporada ao servidor a capacidade de interpretar valores de determinados campos, armazenando o resultado numa tabela para que estes possam ser imediatamente fornecidos caso solicitados em uma requisição DiGIR.

O componente que atua como cliente é um aplicativo desenvolvido em Java, por nós denominado spLinker, descrito a seguir.

3.4.4. SPLINKER

A utilização de um Mirror como provedor equaciona o problema de conectividade, porém gera o problema de envio dos dados da coleção ao Mirror. Para resolver este problema foi criado o aplicativo spLinker.

Assim, o **spLinker** é um aplicativo desenvolvido pelo CRIA para migrar os dados das coleções que não estejam com uma conexão Internet capaz de satisfazer os quesitos de velocidade e estabilidade necessários à instalação de um servidor web.

Para a concepção do spLinker foram utilizadas as seguintes premissas:

- Ser o mais independente possível de alguma plataforma. Isto porque não sabemos qual a plataforma utilizada por coleções que queiram entrar no projeto futuramente e é uma premissa do projeto impor o mínimo de mudanças e/ou exigências técnicas às coleções.
- Exigir a instalação do menor número possível de software auxiliares.

- Ser de fácil entendimento e utilização.

Com base nestas premissas foi escolhido desenvolver o spLinker na linguagem Java. Java é uma das linguagens orientadas a objeto mais independentes de plataforma e sua execução exige apenas a instalação da jvm (Java Virtual Machine) que nas versões mais atuais do Windows já é instalada com o sistema.

A figura a seguir mostra a interface do spLinker.

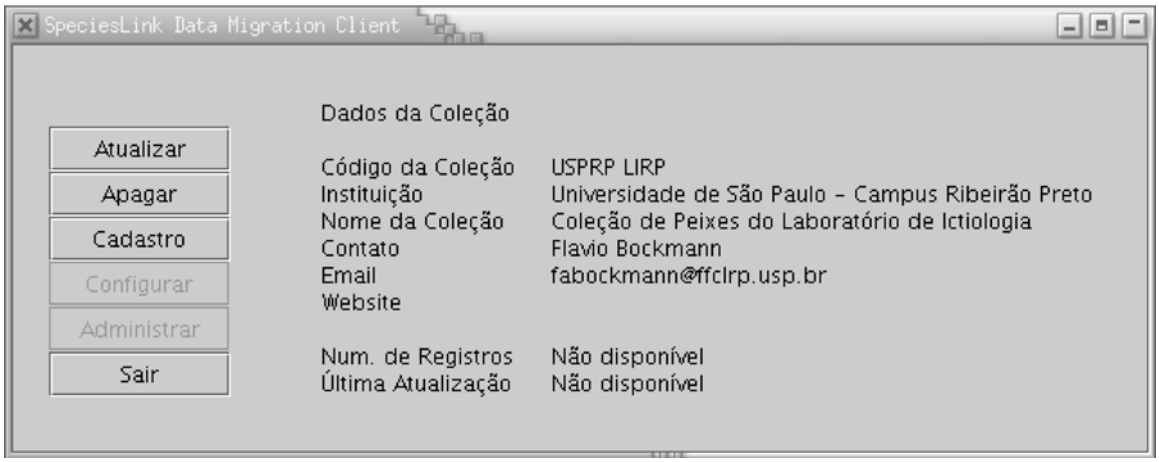


Figura 2. Interface do spLinker

Ao usuário é permitido realizar apenas três operações:

- **Atualizar**: os dados serão lidos da base de dados da coleção, comparados com um repositório local que contém os últimos dados enviados ao Mirror e serão enviados apenas os registros novos (a inserir) ou os removidos (a remover). Os registros modificados são tratados como uma remoção seguida de uma inserção.

  Note que este mecanismo de enviar apenas as alterações da base da coleção, apesar de replicar novamente a base, é fundamental se pensarmos em uma conexão precária da coleção com o Mirror, como no caso de uma conexão por linha discada a 56 Kbps.

- **Apagar**: este comando envia uma requisição ao Mirror pedindo que todos os registros pertencentes à coleção em questão sejam apagados. Em seguida, o repositório local é removido.

- **Cadastro**: lê os metadados cadastrados no Mirror sobre a coleção.

O spLinker foi projetado para ler os dados das coleções através de:

- Conexões aceitas pelo JDBC[19]: O JDBC é a forma padrão para acesso a bancos de dados SQL quando se programa em Java.

- Arquivo texto: caso não seja possível o acesso aos dados via JDBC, basta conseguir gerá-los em arquivo texto que o spLinker será capaz de lê-los e enviá-los ao Mirror. Esta solução foi implementada pois alguns programas

[19] JDBC Java Database Connectivity (java.sun.com/products/jdbc/)

gerenciadores de coleções mantêm os dados em formato proprietário, porém permitem a exportação no formato texto, como é o caso do Biota.

É importante notar que a introdução de novas formas de leitura dos dados é de fácil implementação, pois se utiliza o paradigma de orientação a objetos.

A configuração do spLinker é feita através de arquivos texto. Em sua configuração são preenchidas informações sobre:

- Como se conectar ao mirror (URL e URI).
- Os dados sobre a coleção: identificador, nome, instituição, etc.
- O mapeamento entre os nomes dos campos (JDBC) ou colunas (texto) dos dados originais da coleção e os respectivos nomes utilizados no esquema conceitual utilizado pelo DiGIR (para o speciesLink utiliza-se os campos descritos no DarwinCore).
- O tipo de acesso: JDBC ou Texto
- Informações sobre o acesso via JDBC: URL, driver, usuário e senha da conexão SQL, cláusulas *FROM* e *WHERE* para o comando *SELECT* a ser utilizado para extrair os dados da base da coleção. Note que modificando a cláusula *WHERE* pode-se filtrar dados que por algum motivo não possam ser disponibilizados.
- Informações sobre o acesso via Texto: nome do arquivo, tipo de codificação do arquivo, nome da coluna a ser utilizada como filtro e expressão regular a ser utilizada como filtro.

3.4.5. PORTAL

O portal é o módulo responsável por receber as requisições feitas pelos aplicativos de apresentação dos dados e distribuí-la entre os provedores de dados. Ele também é responsável por juntar as respostas dos vários provedores em uma única resposta e devolver ao aplicativo de apresentação que a requisitou. Pode-se pensar no portal como um centralizador de provedores.

Os comandos aceitos pelo portal são os definidos pelo protocolo DiGIR, já explicado anteriormente.

No contexto do DiGIR, o portal foi concebido de tal forma que ao implementar um aplicativo de apresentação não seja necessário toda vez reimplementar mecanismos necessários à consulta em bases de dados distintas.

Os aplicativos de apresentação não precisam conhecer diretamente os provedores de dados. Eles precisam somente conhecer o endereço do portal e seu protocolo. O portal consegue tratar a complexidade de gerenciar várias requisições em paralelo e unificar a resposta. Um desenvolvimento futuro será o monitoramento dos provedores, mantendo informações tais como: se o provedor está no ar, há quanto tempo está no ar e qual o tempo de resposta dos que estão no ar.

O Portal foi implementado como sendo um Java Servlet através da API padrão descrita no pacote javax.servlet. Para o gerenciamento do servlet que implementa o portal foi escolhido o “servlet container” Tomcat desenvolvido pelo projeto “Apache Jakarta Project” (http: //jakarta.apache.org). O servidor web utilizado para tratar do protocolo HTTP é o Apache (http: //www.apache.org). Estas tecnologias foram escolhidas por serem amplamente utilizadas no mercado e por serem livre (Java) ou open source (Tomcat e Apache).

3.4.6. APLICATIVOS DE APRESENTAÇÃO

Uma das estratégias empregadas no desenvolvimento do DiGIR é a de manter desacoplados os aplicativos que apresentam os dados recuperados a partir da rede do servidor que os recupera. Assim, é possível desenvolver diferentes interfaces para acesso aos dados, que podem oferecer diferentes opções de visualização para os usuários.

No caso do speciesLink foram desenvolvidas duas interfaces para acesso aos dados: uma interface simplificada e outra mais genérica e complexa.

A primeira consiste de uma busca mais simples em que o usuário pode especificar a busca através de alguns campos pré-definidos (Figura 3).

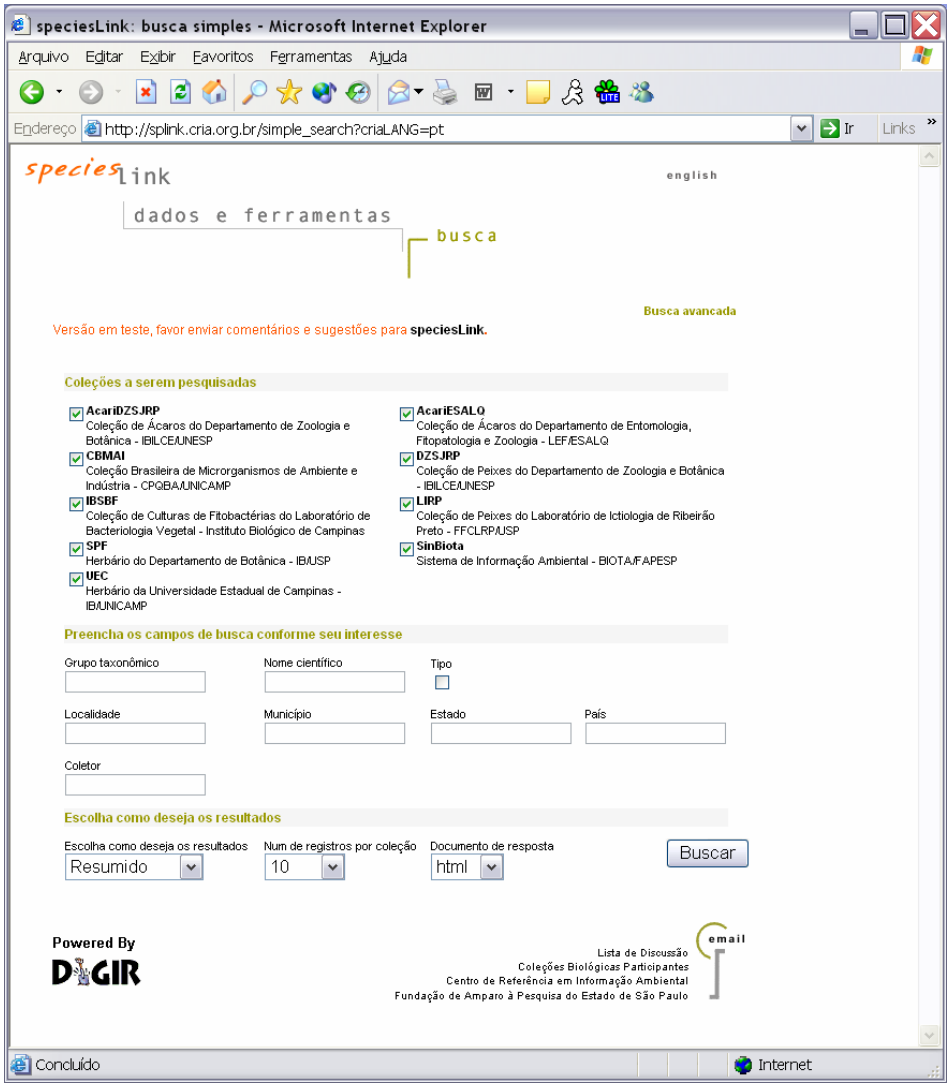


Figura 3. Tela da interface de busca simples do speciesLink

O aplicativo de apresentação dos dados pode ser visto como um cliente DiGIR, ou seja, ele é um aplicativo que pode ser desenvolvido por qualquer um que conheça o protocolo e que queira compartilhar dados disponibilizados por algum **Portal** DiGIR.

Deve-se notar que os aplicativos de apresentação, apesar do nome, não precisam se restringir a apresentar os dados, podendo ser implementados para utilizar os dados de um Portal e gerar análises, por exemplo.

Para o projeto speciesLink foi desenvolvida a biblioteca DiGIR na linguagem Perl que permite a fácil implementação dos aplicativos de apresentação. Esta biblioteca foi integrada ao site do projeto para permitir buscas no Portal via uma interface web mais intuitiva (http://splink.cria.org.br/simple_search) e outra mais genérica (http://splink.cria.org.br/search).

A biblioteca DiGIR implementa métodos relacionados às seguintes necessidades:

- Realizar as consultas previstas pelo protocolo DiGIR: ler os metadados das coleções; realizar buscas nas coleções; pedir o inventário.
- Selecionar: as coleções a serem pesquisadas, o número de registros a retornar, campos a serem retornados, etc.
- Gerar a saída em HTML, permitindo algumas formatações como: mostrar ou não um cabeçalho com o nome dos campos, separar a resposta de cada coleção, etc.
- Ler a resposta em variáveis da linguagem (hashes). Isto é muito útil para a utilização por outros programas.

Atualmente a biblioteca está sendo transformada em um pacote Perl a ser disponibilizado no site oficial do projeto DiGIR.

A figura 4 a seguir procura mostrar todos os elementos apresentados em um diagrama da rede *species*Link.

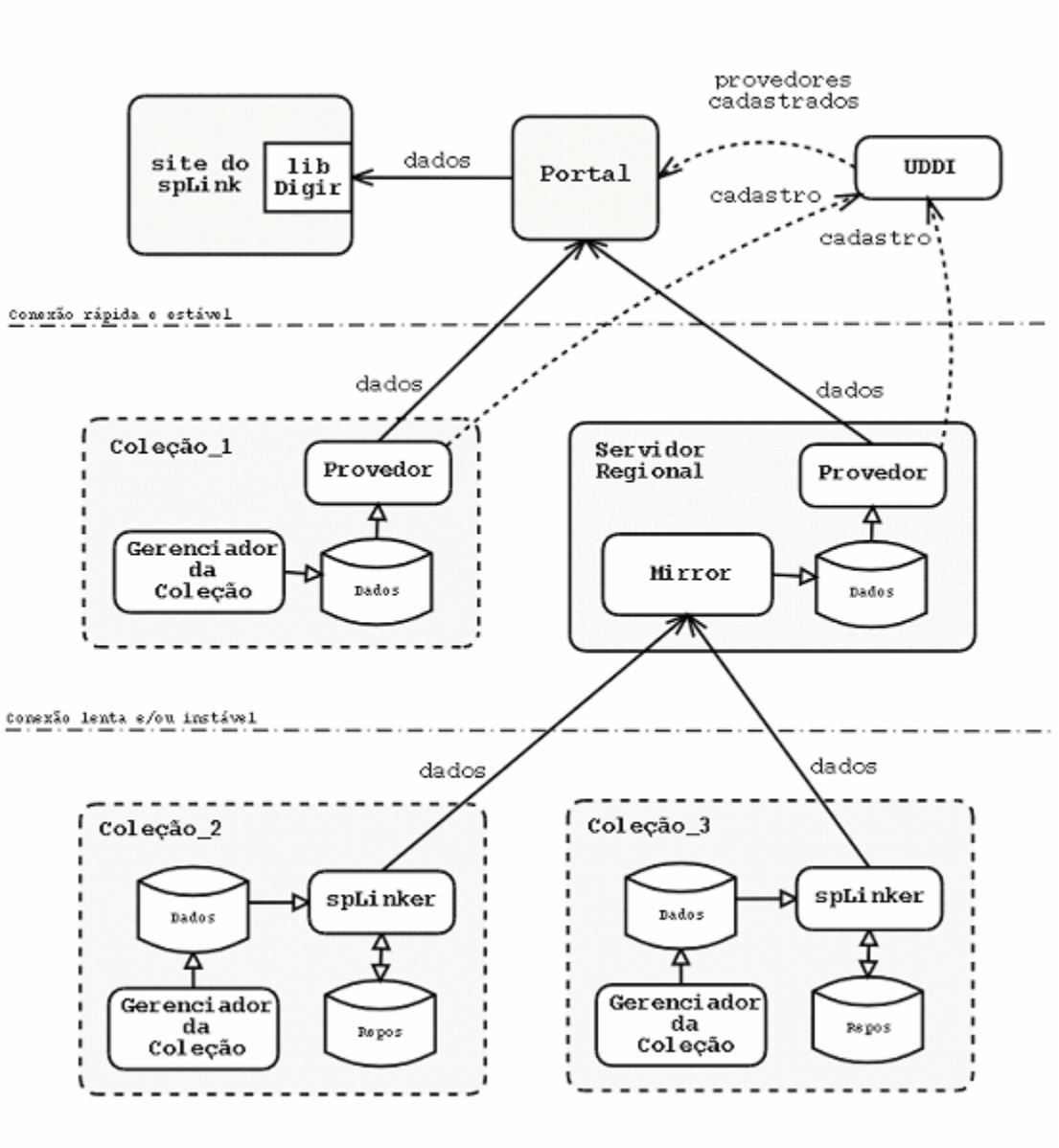


Figura 4. Diagrama da rede do speciesLink

3.4.7. INFRA-ESTRUTURA ADQUIRIDA PELO PROJETO

Módulo Gerenciador da Coleção:

- Quatro microcomputadores Dell Optiplex GX240 com processador Pentium IV e 256 MB de memória RAM (NF 106730) entregues às seguintes coleções:
  - ⇒ Coleção de Algas do Herbário Científico do Estado Maria Eneyda P. Kauffmann Fidalgo, Instituto de Botânica de São Paulo. Representante: Carlos Eduardo de Mattos Bicudo
  - ⇒ Laboratório de Ictiologia de Ribeirão Preto – LIRP, Depto de Biologia – FFCLRP/USP. Representante: Flávio A. Bockmann. CRIA 076/03, 29/05/2003.

⇒ Herbário do Depto Botânica – SPF, Instituto de Biologia da USP. Representante: José Rubens Pirani. CRIA 111/03, 11/06/2003 – (contrato datado de 15/01/2003).

⇒ Coleção de Peixes do Museu de Zoologia – MZUSP – Laboratório de Ictiologia do Depto de Biologia da USP. Representante: Osvaldo Takeshi Oyakawa CRIA 126/03, 18/05/2003.

- Cinco microcomputadores Dell Optiplex GX 240 com processador Pentium IV e 512 MB de memória RAM (NF 106744) entregues às coleções:

  ⇒ Coleção de Acari do Depto de Entomologia, Fitopatologia e Zoologia Agrícola / ESALQ / USP, Piracicaba. Representante: Gilberto José de Moraes.

  ⇒ Coleção de Peixes do Depto de Zoologia e Botânica da Universidade Estadual Paulista – UNESP, São José do Rio Preto. Representante: Francisco Langeani Neto

  ⇒ Coleção de Acari do Depto de Zoologia e Botânica da Universidade Estadual Paulista – UNESP, São José do Rio Preto. Representante: Reinaldo Fazzio Feres

  ⇒ Coleção de Fanerógamas do Herbário do Estado "Maria Eneyda P. Kaufmann Fidalgo" do Instituto de Botânica de São Paulo. Representante: Maria Candida Henrique Mamede

  ⇒ Herbário da Universidade Estadual de Campinas (UEC), Departamento de Botânica, IB, UNICAMP. Representante: Washington Marcondes-Ferreira

- Um microcomputador Dell Optiplex GX 240 com processador Pentium IV e 128MB de memória RAM (NF 083478) entregue à:

  ⇒ Coleção Brasileira de Microorganismos de Ambiente e Indústria – CBMAI / CPQBA / UNICAMP. Representante: Gilson Paulo Manfio. CRIA 165/03 – 08/07/2003 – (contrato datado de 12/03/2002).

- Um microcomputador Dell Precision 330 com processador Pentium IV e 512 MB de memória RAM (NF 013558) entregues à:

  ⇒ Coleção de Peixes do Depto de Zoologia e Botânica (DZSJRP) – IBILCE/ da Universidade Estadual Paulista – UNESP, São José do Rio Preto. Representante: Francisco Langeani Neto. CRIA 049/03, 13/05/2003.

- Um microcomputador Dell Optiplex GX 260 com processador Pentium IV e 512 MB de memória RAM (NF 153639) entregue à coleção:

  ⇒ Herbário do Instituto Agronômico de Campinas (IAC). Representante: Luís Carlos Bernacci. Crono CRIA 131/03, 24/06/2003.

- Dois *no-breaks* PowerWare 5115 1.4 KVA (NF 025.123) instalados nos seguintes locais:

  ⇒ Herbário do Instituto Agronômico de Campinas (IAC). Representante: Luís Carlos Bernacci.

  ⇒ Laboratório de Ictiologia de Ribeirão Preto – LIRP, Depto de Biologia – FFCLRP/USP. Representante: Flávio A. Bockmann

- Quatro câmeras WebCam Creative e três microfones de ouvido (NF 000555) sendo que três estão instaladas nos seguintes locais:

⇒ Coleção de Acari do Depto de Entomologia, Fitopatologia e Zoologia Agrícola / ESALQ / USP, Piracicaba. Representante: Gilberto José de Moraes.

⇒ Coleção de Peixes do Depto de Zoologia e Botânica da Universidade Estadual Paulista – UNESP, São José do Rio Preto. Representante: Francisco Langeani Neto

⇒ Coleção de Acari do Depto de Zoologia e Botânica da Universidade Estadual Paulista – UNESP, São José do Rio Preto. Representante: Reinaldo Fazzio Feres

- Doze licenças do software Microsof Office XP (NF 000506; NF 000334; NF 24253

Módulo servidores regionais

- Três servidores PowerEdge 1400SC (dual processor) com processador Pentium III 933 MHz e sistema operacional Microsoft Windows 2000 Server (NF 075796 e NF 071060) instalados em Campinas, São José do Rio Preto e São Paulo

Módulo Portal

- Dois servidores Dell PowerEdge 6600, cada um com 4 processadores Intel Pentium III Xeon, sistema operacional RedHat Linux 8.0, 2GB de memória RAM por processador e capacidade de disco de 380 GB por servidor. (NF 127902)
- Um sistema de armazenagem de dados em fitas (*backup*) Dell/EMC^2 PowerVault 120T. (NF 019414)
- Equipamento de proteção elétrica (*no-break*) Prestige 6000 6kVA. (NF 020717)
- Ar Condicionado AirSplit 24000BTU (NF 2710)
- Hub Encore 16 portas 10/100 Mbps. (NF 166)
- Software para backup em fita e controlador de unidade de armazenagem ArcServe 7 for Linux. (NF 4205 e 4206)

Conexão do CRIA com a Internet2

- *Switch-router* FoundryNet modelo BigIron 3000 importado diretamente pela Fapesp.

Infra-estrutura da equipe de desenvolvimento do projeto

- Seis Notebooks Dell modelo Latitude C610, com processadores Pentium III, 512 MB de memória RAM e sistema operacional Windows 2000 Professional (NF 013530 e NF 014363)
- Três estações de trabalho Dell Precision 330, com processador Pentium IV, 512 MB de memória RAM e sistema operacional Windows 2000 Professional e monitores adicionais de 17” para cada uma delas. (NF 013558 e NF 071030)
- Um microcomputador Dell Optiplex GX240 com processador Pentium IV e 256 MB de memória RAM (NF 092840)
- Máquina Fotocopiadora EP1031 Minolta. (NF 10408M)
- Impressora LaserJet HP 1200N e *printer server* JetDirect 170X. (NF 0141501)
- Três Teclados e quatro Mouses ópticos para uso com os notebooks. (NF 001441)
- Três Monitores de 17” para uso com os notebooks. (NF 000214)

- Duas Webcam Plus USB. (NF 447)
- Scanner HP5490C. (NF 000507)
- Um *hard disk* de 80 GB. (NF NF 000555)
- Software ArcView 8.2 e ArcGIS 8.2 (licença LabKit) para modelagem de dados e edição de mapas. (25 licenças de cada) (NF 1261)
- Software XManager Standard 5-user Pack para acesso remoto a servidores. (ID 940, de 24/01/2002 - Order n.º RE 6508184 – fornecedor DigiBuy
- Doze licenças do software Microsoft Office XP. (NF 24253 (parte); NF 12399 e NF 25133)
- Pacote de licenças do software Active Vírus Defense (anti-vírus). (NF 007636)
- Software EndNote para tratamento de referências bibliográficas. (NF 11445)
- Software Corel Draw 10. (NF 572)

3.5. INTEGRAÇÃO COM OUTROS SISTEMAS: *SIN*BIOTA, *SPECIESANALYST* E OUTRAS REDES DE COLEÇÕES CIENTÍFICAS

3.5.1. *SIN*BIOTA

A integração do *Sin*Biota, o Sistema de Informação Ambiental do Programa BIOTA/FAPESP e o speciesLink foi feita através de uma tabela no banco de dados do primeiro, específica para esta finalidade, que foi conectada diretamente a um provedor DiGIR.

Esta conexão resultou na integração de cerca de 34 mil ocorrências de espécies contidas atualmente no *Sin*Biota que agora se encontram disponíveis para consulta no *species*Link. Estes dados podem agora ser integrados aos dados das coleções em tempo real, e podem ser obtidos em formatos de fácil utilização (tabelas ou planilhas de cálculo).

O *Sin*Biota é o sistema responsável por integrar as informações geradas pelos pesquisadores vinculados ao programa, portanto os dados aqui expostos se restringem às coletas realizadas no âmbito do programa BIOTA. Vale salientar que os dados são fornecidos “como estão” e os mantenedores do sistema não se responsabilizam pela acurácia, confiabilidade e completude dos dados fornecidos.

A tabela criada no *Sin*Biota para fazer a conexão é atualizada automaticamente de hora em hora, portando a inclusão de uma nova coleta no banco de dados do SinBiota é disponibilizada praticamente de forma instantânea no speciesLink

3.5.2. INTEGRAÇÃO COM *SPECIESANALYST* E OUTRAS REDES DE COLEÇÕES CIENTÍFICAS INTERNACIONAIS

Uma vez que o speciesLink utiliza o mesmo protocolo de comunicação que várias outras iniciativas internacionais, é possível integrar dados da rede paulista com fontes de dados de outras redes regionais ou internacionais.

Esta integração se dá em duas direções. A primeira delas é a integração dos dados de coleções internacionais, tornando-as visíveis nas buscas do speciesLink.

Grande parte das coleções internacionais participantes do projeto *SpeciesAnalyst*, assim como novas fontes de dados no Canadá e Europa estão neste momento

migrando para o protocolo DiGIR e deverão estar disponíveis como fontes de dados no speciesLink dentro de pouco tempo.

A primeira fonte de dados que já está disponível para consultas usando o protocolo DiGIR é o FishBase, que deverá ser integrado até o final do mês de junho à rede do speciesLink. Esta fonte de dados servirá como base para os testes de integração com fontes de dados internacionais, uma vez que contém dados complementares às três coleções de peixes paulistas que participam do speciesLink, sendo que duas delas já se encontram conectadas ao sistema.

De maneira simétrica, as coleções paulistas podem ser conectadas aos portais das redes internacionais de maneira simples e automática.

3.5.3. INTEGRAÇÃO COM OUTROS SISTEMAS

O CRIA trabalha com sistemas de informação nos quais o nome científico de organismos é o objeto principal. É o caso do speciesLink, do *Sin*Biota, do SICol, do Neofrug. Trabalhamos também com sistemas onde os nomes científicos, apesar de não serem o objeto central, são citados e representam uma rica fonte de informação complementar, como é o caso da revista Biota Neotropica, do Bioline International e Imagens da Biodiversidade Brasileira.

Foi através da constatação desse fato que se decidiu criar mecanismos de indexação e recuperação de informação que permitissem o cruzamento das informações relativas ou associadas aos organismos existentes nos diferentes sistemas.

Para aqueles que têm uma estrutura de dados própria, como é o caso do *Sin*Biota ou SICol, foram criados servidores SOAP específicos, escritos em Perl que respondem a requisições que têm como objeto um nome científico e como resposta a existência de informação relativa a ele no sistema.

Para os outros sistemas, mormente os textuais, como é o caso dos artigos da revista Biota Neotropica ou existentes no Bioline International, um banco de dados centralizado foi implementado para tratar da ocorrência de informação relativa aos nomes científicos. Também nesse caso, um servidor SOAP (*xtaxa*) é responsável pelas requisições.

Existem ainda alguns bancos de dados menores que utilizam um aplicativo genérico, com modelo cliente/servidor SOAP, desenvolvido no CRIA chamado xinfo, que é capaz de tratar informação textual que possa ser organizada em registros. Também esse aplicativo interage com o xtaxa indexando qualquer nome científico que ocorra nos textos por ele tratados.

Dessa forma, qualquer citação de um nome científico, em qualquer página de informação gerada por qualquer sistema mantido pelo CRIA, é associada a uma imagem 

. Um link é criado dinamicamente de modo que, quando a imagem é clicada, uma nova página é apresentada ao usuário com o resultado da busca pela espécie a ela associada em todos os sistemas do CRIA. A figura a seguir mostra um exemplo de busca integrada.

Figura 5. Página integrando a busca pelo nome da espécie, no caso Hoplias

Nessa mesma página, alguns sistemas externos como SCiELO, PubMed, etc., são apresentados e a ocorrência de informação associada à espécie é verificada.

Além dos índices já apresentados, o CRIA mantém uma cópia resumida do banco de dados do Catálogo da Vida do *Species 2000*, contendo cerca de 470.000 nomes de

espécies. Este banco de dados serve de parâmetro para comparação entre a espécie buscada e o nome válido, fornecendo quando possível informações taxonômicas completas, ou mesmo sugerindo nomes semelhantes e listas de espécies e infra-espécies.

3.6. REPATRIAÇÃO DE DADOS

Durante o período de 14 de Janeiro a 01 de Março de 2002 foram realizadas visitas ao Royal Botanical Garden of Edinburgh – RBGE e ao Kew Gardens – Londres/Inglaterra pela pesquisadora do CRIA Marinez Ferreira de Siqueira.

O objetivo foi levantar informações que servissem de base para avaliar o trabalho necessário para a repatriação das informações sobre tipos de espécies brasileiras depositados nesses herbários. O objetivo era responder às seguintes questões: Quais as famílias deverão ser contempladas? Qual a quantidade de trabalho envolvido para cada família? Qual ou quais ecossistemas brasileiros esses herbários melhor representam?

No caso do herbário do Jardim Botânico de Edimburgo, foi também feita uma experiência de digitalização de exsicatas tipo de espécies do cerrado brasileiro. O relatório desta viagem encontra-se no anexo 1.

Quanto ao Kew Gardens, este herbário já havia iniciado um projeto de repatriação de dados "Repatriation of Herbarium Data project for Northeastern Brazil[20]" em parceria com a Associação Plantas do Nordeste.

3.7. DESENVOLVIMENTO DE APLICATIVOS: MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES

O nível de conhecimento sobre a distribuição geográfica de espécies tropicais é precário. A maioria das espécies é representada por poucos pontos de amostragem e o georeferenciamento, principalmente dos dados históricos disponíveis nas coleções biológicas, é muito impreciso.

O programa Biota/Fapesp introduziu a obrigatoriedade do uso do GPS (Global Positioning System) nas pesquisas de campo realizadas no âmbito do programa e tornou obrigatório o compartilhamento do resultado das pesquisas. O *Sin*Biota[21] foi desenvolvido para armazenar e disponibilizar esses dados juntamente com a base cartográfica do Estado de São Paulo (produzida pelo Instituto Florestal). Por sua vez o projeto speciesLink está desenvolvendo o catálogo virtual das coleções científicas do Estado de São Paulo.

A existência dessa nova base de informação a serviço da pesquisa, torna factível estudar e desenvolver ferramentas que possam usar esta infra-estrutura para produzir informações para análise e tomada de decisão.

A equipe do CRIA começou a se envolver com modelagem a partir da vinda do Prof. Townsend Peterson ao Brasil em 1999 e a ida de um membro da equipe à Universidade de Kansas para trabalhar no desenvolvimento do GARP (Genetic Algorithm for Rule-set Production). A aprovação do projeto speciesLink em outubro de 2001 deu ao CRIA a oportunidade e as condições necessárias para desenvolver esta linha de pesquisa no Brasil.

[20] http://www.kew.org/data/repatbr/about.html
[21] Sistema de Informação Biota/Fapesp: http://sinbiota.cria.org.br

O desenvolvimento dessa linha de pesquisa no escopo do projeto speciesLink tem por objetivo estudar, desenvolver e disseminar ferramentas de modelagem de distribuição geográfica adequadas para as espécies e para as condições ecológicas/ambientais brasileiras.

Trata-se de um trabalho que necessariamente envolve o estabelecimento de parcerias tanto no desenvolvimento (Universidade de Kansas) como no uso, teste e aplicação das ferramentas (pesquisadores de instituições do Brasil e do exterior).

O processo de modelagem de nicho ecológico consiste em converter dados primários de ocorrência de espécies em mapas de distribuição geográfica indicando a provável presença ou ausência da espécie, neste caso, através da aplicação de algoritmo genético.

Estes modelos trabalham, na maioria dos casos, com o conceito de nicho ecológico fundamental da espécie. Este conceito foi definido por MacArthur (1972), como sendo um conjunto de condições ecológicas com as quais as populações conseguem se manter, que pode ser representado por um espaço ecológico/ambiental multidimensional (Figura 6).

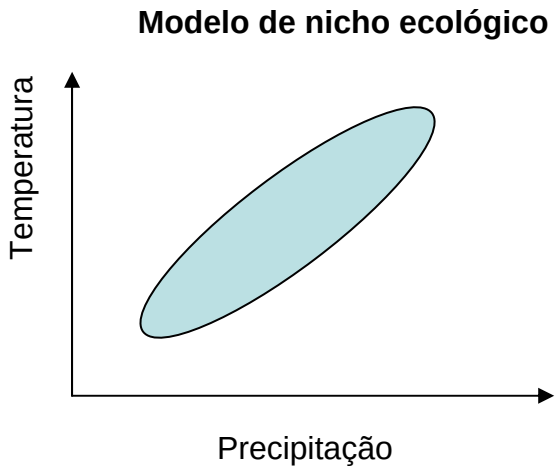


Figura 6. Exemplo de espaço bidimensional definido como nicho ecológico de espécie baseado em dois parâmetros ambientais (temperatura e precipitação).

Tais algoritmos tentam encontrar relações não-aleatórias entre os dados de ocorrência da espécie com os dados ecológico/ambientais relevantes para a espécie (tais como: temperatura, precipitação, topografia, tipo de solo, geologia, entre outros) no ponto onde a espécie foi registrada (Figuras 7 e 8).

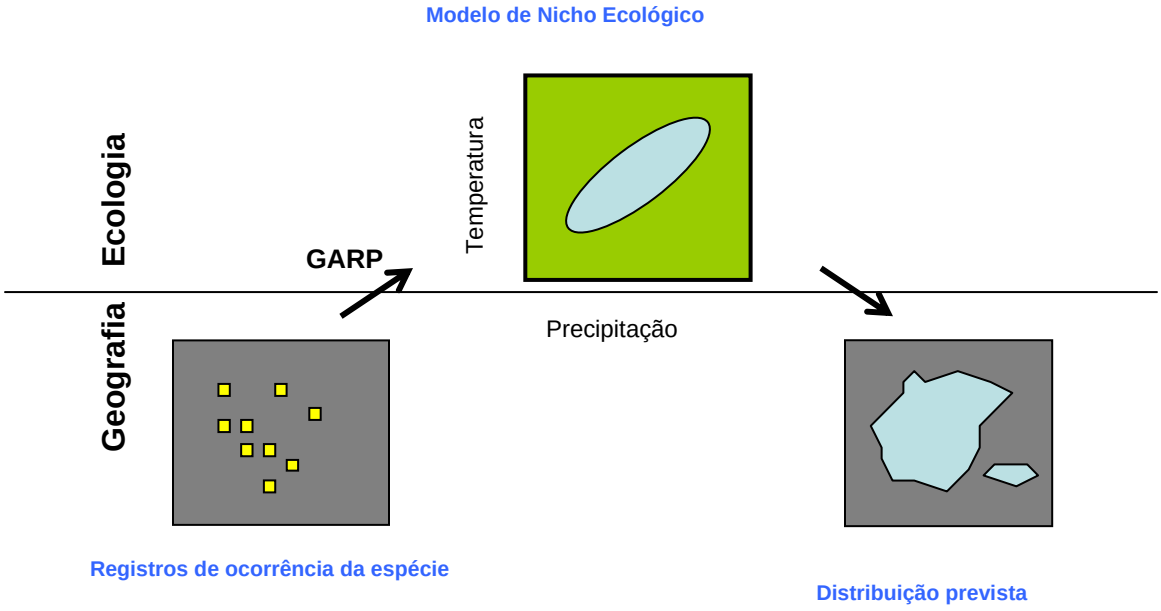


Figura 7. Esquema mostrando a relação entre os dados de campo e a previsão de distribuição geográfica através do processo de modelagem de nicho ecológico.

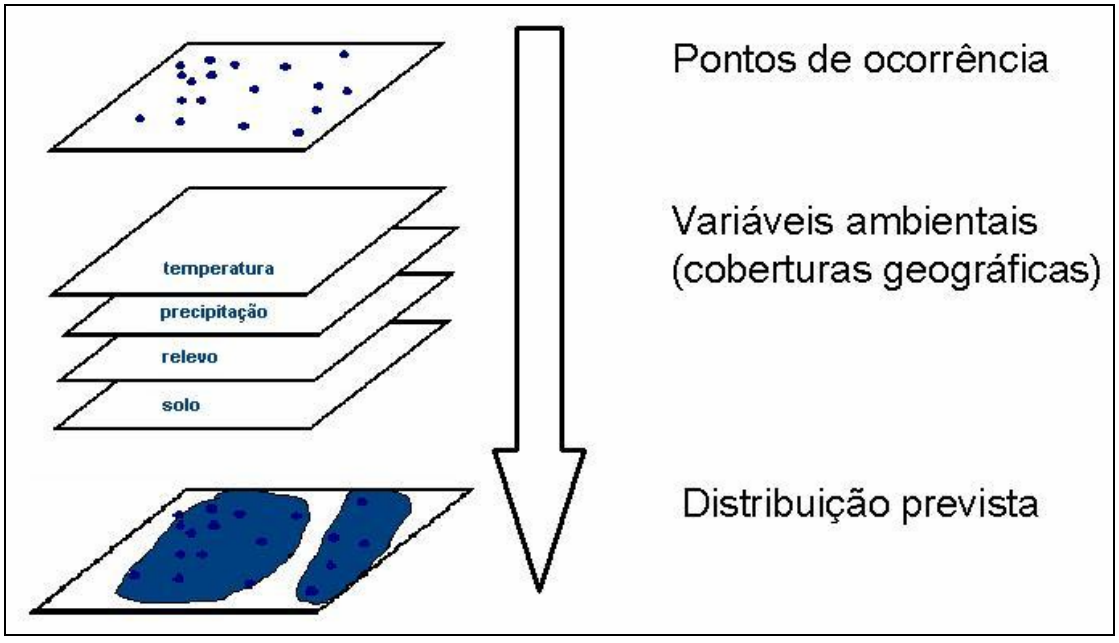


Figura 8. Esquema mostrando a relação entre as variáveis ambientais e os pontos de ocorrência da espécie.

Existem vários métodos na literatura que podem ser aplicados para se obter modelos de nicho ecológico a partir dos pontos de ocorrência de espécies. Apresentamos no anexo 2 um relatório completo da pesquisa realizada desde a vinda do Prof. Townsend Peterson em maio de 2002 até junho de 2003. Os estudos incluíram:

- Estudos sobre algoritmos genéticos:
  - ⇒ Inspiração na Teoria da Evolução das Espécies e Terminologia
  - ⇒ Espaço de Busca de Soluções e as Superfícies de Otimização
  - ⇒ Operadores Heurísticos
  - ⇒ Genetic Algorithm for Rule-set Production (GARP)
- Modelo de Nicho Ecológico no GARP
- Estudo e Análise dos Dados de Biodiversidade
- O desenvolvimento de vários projetos como:
  - ⇒ Modelagem de espécies de distribuição geográfica restrita baseada em similaridade ambiental. Marinez Ferreira de Siqueira (CRIA), Giselda Durigan (IF-Assis), Ricardo S.Pereira (CRIA) e A. Townsend Peterson (U. Kansas).
  - ⇒ Modelagem de espécies de plantas arbóreas da Bacia do Médio Paranapanema - Marinez Ferreira de Siqueira (CRIA), Giselda Durigan (IF-Assis), Wilson Aparecido Contieri (IF-Assis), Ricardo S. Pereira (CRIA) e A. Townsend Peterson (U. Kansas).
  - ⇒ Determinação de áreas para estudos de controle biológico do ácaro vermelho do tomate - Rafael Luís Fonseca (CRIA), Ricardo Scachetti Pereira (CRIA), Imeuda Peixoto Furtado (ESALQ) , Dr. Gilberto José de Moraes (ESALQ)
  - ⇒ Modelos de nicho potencial de espécies vegetais dispersas pela megafauna: os efeitos da perda de dispersores. Rafael Luís Fonseca (CRIA), Dr. Mauro Galetti (UNESP – Rio Claro)
  - ⇒ Sensibilidade do GARP a dimensionalidade dos dados: um teste usando análise de componentes principais (PCA). Rafael Luís Fonseca (CRIA), Ricardo Scachetti Pereira (CRIA), Rafael Luís Galdini Raimundo (Unicamp), Dr. Thomas M. Lewinsohn (Unicamp)
  - ⇒ Distribuição geográfica de duas espécies crípticas de Tomoplagia (Diptera: Tephritidae): condicionantes em macroescala. Aluana Gonçalves de Abreu (Unicamp), Marcio Uehara-Prado (Unicamp) e Rafael Luís Fonseca (CRIA)

Como produtos temos a aplicação do GARP no Brasil para prever :

- a distribuição geográfica de espécies;
- algumas conseqüências de alterações climáticas futuras na distribuição geográfica de espécies; e
- o potencial de invasão de espécies exóticas.

Um relatório completo sobre a pesquisa realizada encontra-se no anexo 2 e é parte integrante deste relatório.

3.8. OUTROS DESENVOLVIMENTOS

3.8.1. MAPSERVER/MAPSCRIPT

O projeto *Sin*Biota teve como um de seus objetivos o estudo e implementação de uma ferramenta gráfica de visualização da distribuição de espécies coletadas utilizando os

dados disponíveis no sistema e a base cartográfica em escala 1:50.000 especialmente desenvolvida pelo Instituto Florestal para o projeto.

No início do projeto foi definido o uso de tecnologias proprietárias como o ArcView, ArcIMS, além de gerenciador de banco de dados Oracle, para garantir o sucesso da implementação desse objetivo. Porém com o tempo, essas alternativas foram se mostrando, além de economicamente inviáveis, fontes de muitos problemas de implementação em diferentes plataformas, além de representarem uma enorme dependência das empresas proprietárias. Um dos problemas mais marcantes que encontramos foi o alto custo das licenças necessárias do ArcView para que fosse possível a disponibilização da interface na Internet.

Ainda no escopo do projeto *Sin*Biota foram buscadas soluções alternativas para o problema. Decidiu-se pelo desenvolvimento de uma aplicação de apresentação em Perl que utilizava alguns componentes, também escritos em Perl, capazes de se comunicar com o MapObjects, *software* também proprietário, porém muito mais barato e com a possibilidade de uso de até 50 instâncias em ambiente MS-Windows. Dessa forma, a aplicação web resolvia todas as questões específicas do sistema e, através de sockets, delegava aos servidores de mapa, em ambiente MS-Windows, a função de desenhar o mapa desejado e devolvê-lo à aplicação web.

Essa arquitetura funcionou bem durante muito tempo. Porém, com a inclusão de novos *layers*, em alguns casos o processo passou a ter resposta bastante lenta. Um exemplo é o *layer* sobre a drenagem do estado que demorava cerca de 40 segundos para ser desenhada.

A integração *Sin*Biota ↔ speciesLink pressupõe a utilização plena de todas as ferramentas disponíveis nos dois sistemas. É objetivo da equipe trabalhar com interfaces gráficas e utilizar toda a base cartográfica disponível para o Estado de São Paulo também para o speciesLink. A nova demanda do projeto, aliada aos problemas de performance do modelo em uso, fez com que a equipe estabelecesse um grupo para estudar novas opções para o problema com vistas a apresentar uma solução que pudesse ser utilizada no contexto do speciesLink.

Após a análise de algumas possibilidades, foi decido investir no desenvolvimento de um módulo específico que tivesse como base a biblioteca MapScript do *software* conhecido como MapServer. O MapServer foi originalmente desenvolvido pelo projeto ForNet da Universidade de Minnesota (UMN) em cooperação com a NASA e o Departamento de Recursos Naturais de Minnesota (MNDNR). Melhorias foram posteriormente feitas pelo MNDNR e o Minnesota Land Management Information Center (LMIC). Os desenvolvimentos atuais são financiados pelo projeto TerraSIP, financiado pela NASA, entre a UMN e um consórcio de interesse em gerenciamento de terra.

O projeto é desenvolvido como *open source* e utiliza outros módulos também *open source* ou livres. Utilizando a biblioteca MapScript, foi então desenvolvido um módulo genérico, escrito em linguagem Perl, capaz de substituir os servidores de mapas em MS-Windows usando o MapObjects. Esse módulo foi desenvolvido no escopo do projeto e está sendo aprimorado para que se torne um componente genérico a ser utilizado por outras aplicações que necessitam da geração de mapas, independentemente do escopo em que a aplicação está sendo executada.

O ganho com esse novo módulo foi enorme uma vez que dispensou o uso de máquinas auxiliares em ambiente MS-Windows, permitindo que todo o processo seja executado

em um único passo, além de melhorar a performance em cerca de 10 vezes em relação à solução anterior.

3.8.2. IMAGENS DA BIODIVERSIDADE BRASILEIRA: O DESENVOLVIMENTO DE UM PROTÓTIPO

Através do contato direto com pesquisadores envolvidos no projeto speciesLink, *Sin*Biota e outros, foi possível constatar que grande parte deles possui acervos importantes de fotografias associadas a seus trabalhos de pesquisa. De modo geral, os pesquisadores não dispõem de mecanismos eficientes que lhes permitam partilhar essas fotografias com outros colegas ou com a comunidade em geral. Assim, geralmente, uma pequena parte dessas fotos é utilizada em publicações científicas e o restante acaba ficando indisponível. Associado a isso, percebemos também o interesse das coleções científicas de, na digitalização de seus acervos, incluírem imagens das exsicatas, principalmente dos tipos.

O CRIA tem por objetivo disponibilizar informação sobre a biodiversidade brasileira, principalmente tendo como foco o nome científico das espécies, agregando a ele o maior número possível de informações complementares. Sendo assim, o CRIA decidiu desenvolver um web site específico que permita aos pesquisadores armazenar, qualificar, manter e disponibilizar suas fotos de maneira organizada e autônoma.

Foi desenvolvido um protótipo que já se encontra em teste no endereço: imagem.cria.org.br, que conta hoje com pouco mais de 3000 imagens, sendo que pouco mais de 2000 já estão disponíveis para o público em geral.

O sistema é composto basicamente por dois módulos, ambos escritos em linguagem Perl, utilizando PostgreSQL como servidor de banco de dados e *software* abertos como Imager e ImageMagick para manipulação das imagens.

O primeiro módulo é o de manutenção e é acessado pelos autores das fotos e outros pesquisadores que colaboram com o projeto na identificação das espécies objeto das fotos.

Cada autor é cadastrado no sistema e recebe uma senha que lhe dá acesso às suas fotos e à informação a elas associadas. O autor pode permitir ainda que outros pesquisadores colaboradores tenham acesso aos dados relativos às suas fotos para auxiliarem na identificação das espécies.

Uma interface web foi especialmente desenvolvida para controlar todo o processo.

O segundo módulo desenvolvido é de acesso público e disponibiliza as fotos na Internet através de um formulário específico de busca através do qual o usuário pode realizar buscas combinadas nas informações associadas às fotos. O resultado é apresentado em forma de thumbnails que, quando clicadas, apresentam as fotos em tamanho grande com toda a informação a ela associada.

Quando a informação associada contém o nome de espécies, uma ligação com o sistema de busca por nome de espécies do CRIA é automaticamente inserida de modo a permitir o acesso a todas as informações existentes no CRIA sobre aquela espécie.

Depois de efetuar uma busca, o usuário pode ainda ver um slide show das fotos encontradas.

Um especial cuidado foi tomado desde o início no que se refere à preservação da autoria tanto das fotos quanto da identificação das espécies que nelas aparecem. Para isso, o nome do autor é inserido na foto juntamente com uma marca d'água do CRIA, e o nome do identificador é citado junto à informação associada. Além disso, apesar das

fotos originais ficarem armazenadas no sistema, apenas versões de baixa resolução são disponibilizadas ao público.

3.8.3. QUALIDADE DE DADOS

Estão sendo criadas listas de referências com nomes de espécies, coletores e localidades, além de coordenadas geográficas, para auxiliar no controle da qualidade dos dados no processo de informatização das coleções. A partir destas listas, e dos erros detectados nos bancos de dados das coleções, será possível desenvolver ferramentas para assegurar a qualidade dos dados. Um subproduto deste trabalho é a disponibilização de bancos de dados na WEB. Um exemplo, em fase de implementação, é o "Banco de coletores de plantas do Brasil". O banco está sendo elaborado a partir dos dados da coleção do herbário UEC, somados às abreviaturas dos nomes dos coletores contidos nos volumes 1 e 2 da Flora Fanerogâmica do Estado de São Paulo (Wanderley et al, 2001, 2002) e dados fornecidos pelos herbários do "New York Botanical Garden", do Instituto Agronômico de Campinas e do Instituto de Botânica de São Paulo. As grafias diferentes e as datas de coletas relacionadas à estas grafias são mantidas,. Desta forma, na maioria das vezes é possível distinguir os homônimos. As informações relacionadas às abreviações, como: o nome completo, a área em que atua ou atuou, o período em que existem coletas desta pessoa em determinado herbário, as datas de nascimento e morte, , a origem do coletor, a Instituição em que trabalha ou trabalhou e observações, são procuradas em fontes diversas. Como exemplos temos, o Index Herbariorum[22], o IPNI[23], a página de taxonomistas da UFRJ[24], listas de ex-alunos e alunos dos cursos de pós-graduação das instituições ligadas ao projeto e páginas de Universidades e Instituições de pesquisa. Posteriormente, as dúvidas serão levadas a pessoas com amplo conhecimento da história de suas respectivas instituições. O objetivo do banco de dados virtual dos coletores é o de reunir o maior número de informações disponível em um primeiro momento e buscar a participação dinâmica e continada dos pesquisadores. Desta forma, o banco poderá auxiliar no controle de qualidade dos dados de coleções de herbário no Species Link, auxiliar o trabalho de taxonomistas e ser um repositório de dados históricos de coletores de Plantas.

A página foi criada utilizando o programa xInfo - Sistema Genérico de Indexação e Recuperação de Dados em XML (*software* próprio desenvolvido pelo CRIA em 2001), que é um sistema de armazenamento e recuperação de dados, basicamente textuais, organizados em XML, de acordo com a DTD definida em xinfo.dtd, utilizando a técnica de cliente/servidor. A idéia deste *software* é a de permitir ao usuário controle sobre seus dados e sua manutenção em um ambiente conhecido (MS-Excel, MS-Access) e dar a ele ferramentas para exportá-los para um sistema de simples utilização capaz de disponibilizar esses dados via web.

3.9. DIFUSÃO

3.9.1. CURSOS, EVENTOS E PALESTRAS

O principal evento realizado no período foi o Fórum internacional "Trends and Developments in Biodiversity Informatics" realizado em Indaiatuba em outubro de 2002.

[22] www.nybg.org/bsci/ih/searchih.html, 06/02/2003
[23] www.ipni.org/ipni/query_author.html, 06/02/2003
[24] www8.ufrgs.br/taxonomia, 06/02/2003

A equipe trabalhou na disseminação do projeto proferindo palestras e participando de reuniões técnicas e o *website* do fórum inclui os resumos e as apresentações de todos os participantes (www.cria.org.br/eventos/tdbi/).

Além do Fórum, a equipe proferiu as seguintes palestras sobre modelagem:

- Workshop – Modelagem de Biodiversidade. Local: Belém Data: 10-12 de Fevereiro de 2003. GEOMA - Rede Temática de Pesquisa em Modelagem Ambiental da Amazônia - MCT. Participantes do CRIA: Ricardo Scachetti Pereira e Marinez Ferreira de Siqueira. Apresentação da palestra: Algoritmos genéticos - GARP
- A Informática como Ferramenta para Conservação da Biodiversidade. II Semana da Biologia UFSCar - 24 a 29 de março de 2003. Universidade Federal de São Carlos – São Carlos – SP. Apresentação: Rafael Luís Fonseca
- Modelagem Preditiva de Distribuição de Espécies. I Curso de Introdução ao Sistema de Informação do Programa Biota/Fapesp (*Sin*Biota). 13 e 14 de fevereiro de 2003 - Instituto de Biologia – Unicamp. Apresentação: Rafael Luís Fonseca
- Apresentação dos trabalhos desenvolvidos no CRIA pela Dora Canhos no workshop "Coleções Biológicas: Desafios e Perspectivas no Brasil", de 5 a 7 de novembro no Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Escola Nacional de Botânica Tropical.
- Modelagem do potencial invasivo de *Lantana camara* L. (Verbenaceae). Apresentação oral no III Segundo Simpósio do Programa Biota/Fapesp. 26 a 28 de novembro de 2002 – Universidade Federal de São Carlos – SP. Apresentação: Rafael Luís Fonseca e Sérgio Rodrigues Morbiolo
- Modelagem Preditiva de Distribuição Geográfica de Espécies. Participação em Aula de Biogeografia – Prof. Dr. Thomas Michael Lewinsohn. Curso de Geografia. 06 de novembro de 2002 – Instituo de Biologia – Unicamp. Apresentação: Rafael Luís Fonseca
- Modelagem Preditiva de Distribuição Geográfica de Espécies. Participação em Aula de Ecologia Básica - Prof. Dr. Paulo Sergio Oliveira. Curso de Ciências Biológicas. 18 de outubro de 2002 – Instituto de Biologia – Unicamp. Apresentação: Rafael Luís Fonseca

- Apresentação dos trabalhos desenvolvidos no CRIA pela Dora Canhos no Painel: Coleções e Sistemas de Informação para Biodiversidade no Encontro da Rede de Inventário, Coleta e Cultivo. Seminário sobre Biodiversidade, Biotecnologia e Bionegócios na Amazônia. 1a. Feira Internacional da Amazônia. Manaus, 12 de setembro de 2002.

- 53o. Congresso Nacional de Botânica. Seminário sobre Bancos de Dados Botânicos no Brasil. Dora Ann Lange Canhos. 23 de julho de 2002. Apresentação e Resumo

3.9.2. PUBLICAÇÕES

Ainda como produtos temos as seguintes publicações:

- Modelagem GARP da distribuição nativa e exótica da planta invasora *Chromolaena odorata* (Asteraceae). Rafael L. G. Raimundo (Unicamp), Rafael L.

Fonseca (CRIA), Ricardo S. Pereira (CRIA), A. T. Peterson (U. Kansas), Thomas M. Lewinsohn (Unicamp). Submetido para "Journal of Applied Ecology".

- Predicting the potential of invasion of two *Crotalaria* species (Fabaceae) in Conservation Units in Brazil. Rafael Luís Fonseca (CRIA), Paulo Guimarães Jr. (Unicamp), Sérgio R. Morbiolo (Herbário Unicamp), Ricardo Scachetti Pereira (CRIA), Townsend Peterson (KUNHM-BRC). Manuscrito finalizado.
- Modelagem do potencial invasivo de Lantana camara l. (Verbenaceae) em ecorregiões e unidades de conservação tropicais. Sérgio R. Morbiolo (Herbário Unicamp) e Rafael Luís Fonseca (CRIA). Submetido para "Biota Neotropica".
- Characterizing geographic distributions of tropical woody plant species via ecological niche modeling. Ingrid Koch (Unicamp), Marinez Ferreira de Siqueira (CRIA), A. T. Peterson (U. Kansas). Status da publicação: Submetido para " Global Ecology and Biogeography" .
- Global Climate Change Consequences for Cerrado Tree Species Distribution. Marinez Ferreira de Siqueira (CRIA) & A. T. Peterson (U. Kansas). Submetido para "Biota Neotropica".
- Avaliação do potencial de invasão de Homalodisca coagulata na Califórnia e no Brasil. A. T. Peterson (U. Kansas), Ricardo Scachetti Pereira (CRIA), Daniel A. Kluza (U. Kansas). Biota Neotrópica 3(1): http://www.biotaneotropica.org.br/v3n1/pt/abstract?article+BN00703012003
- Detectando problemas de identificação em conjuntos de dados sobre biodiversidade baseado em modelagem de nichos ecológicos. A. T. Peterson (U. Kansas), Ingrid Koch (Unicamp), Ricardo Scachetti Pereira (CRIA), Adolfo G. Navarro-Sigüenza (UNAM, Mexico). Submetido para “Diversity and Distributions”.
- Detection of errors in biodiversity data: Collectors’ itineraries flag mislabeled specimens. A. T. Peterson (U. Kansas), Adolfo G. Navarro-Sigüenza (UNAM, México), Ricardo Scachetti Pereira (CRIA, Brasil). No prelo. “Bulletin of the British Ornithologists’ Club”.
- Distribuição de vetores da Leishmaniose cutânea em São Paulo. Vera Camargo Neves (SUCEN), A. Townsend Peterson (U. Kansas), Ricardo Scachetti Pereira (CRIA). Submetido para “Sociedade Brasileira de Medicina Tropical”.
- Distribuição da capivara na bacia do Rio Piracicaba no Estado de São Paulo. Kátia Ferraz (ESALQ, Piracicaba), Ricardo Scachetti Pereira (CRIA), A. T. Peterson (U. Kansas). Aceito para publicação nos anais do IALE2003 (Simpósio sobre landscape ecology).

3.10. BOLSAS IMPLEMENTADAS NO DECORRER DO PROJETO

Bolsa de Pós-doutorado

- Ingrid Koch

Bolsas de Pesquisador Visitante

- Andrew Townsend Peterson
- Arthur Chapman

Bolsas de treinamento técnico – nível 3

- Sérgio Morbiolo (UEC)
- Ana Paula Fortuna (UEC)

- Janete Moscardi (UEC)
- Kátia Freire da Silva (SP)

4. CONCLUSÕES, RECOMENDAÇÕES E DIRETRIZES FUTURAS

Os resultados práticos do projeto estão acima das expectativas da equipe. O desenvolvimento de um protocolo de comunicação como parte de uma iniciativa internacional colaborativa deu ao projeto um destaque internacional o que permitiu avançar, mais do que o esperado no campo da padronização de dados biológicos. Esse fato certamente irá facilitar a integração desse sistema (*species*Link) com outros sistemas internacionais (em especial o *Species Analyst*).

Percebe-se também uma mudança cultural significativa por parte das coleções científicas que hoje têm uma postura muito mais aberta em relação à informatização dos acervos e ao compartilhamento de dados.

Ao longo do projeto percebemos que:

- Não podemos ficar à margem do processo de informatização das coleções, como era a nossa intenção inicial. As coleções precisam de apoio quanto à escolha e uso do *software*, mas é importante frisar que a decisão final sobre a escolha do *software* a ser adotado deve ser da própria coleção.
- O sistema deve ser capaz de se adequar a qualquer *software* que a coleção adote, desde que preencha alguns requisitos mínimos.
- É importante desenvolver ferramentas que auxiliem a coleção na revisão e na melhoria dos dados de seu acervo. Destacamos ferramentas de georeferenciamento automático (proposta sob análise processo 02/08379-4) e ferramentas de *data cleaning* (objeto do pedido de prorrogação do projeto).
- É importante desenvolver ferramentas que possibilitam a comparação dos acervos e uma possível análise de temas como: esforço de coletas, grupos taxonômicos mais e menos coletados; áreas pouco coletadas, etc (objeto do pedido de prorrogação do projeto).

Será solicitada uma prorrogação do prazo do projeto visando:

- A inclusão de novas coleções na rede speciesLink;
- Dar continuidade ao apoio às coleções na estruturação de seus sistemas de informação;
- O desenvolvimento de *guidelines* e ferramentas voltadas para a "limpeza" dos dados (*data cleaning*);
- O desenvolvimento de um aplicativo de visualização dos dados em uma base cartográfica;
- Dar continuidade à linha de pesquisa em modelagem.

5. EQUIPE

Vanderlei Perez Canhos, coordenador

Desenvolvimento:

- Alexandre Marino (banco de dados, SinBiota)
- Fabricio Pavarin (apoio às coleções)
- Mauro Enrique de Souza Munhoz (DiGIR, modelagem)
- Renato de Giovanni (DiGIR)
- Ricardo Scachetti Pereira (apoio às coleções, modelagem, DiGIR)
- Rosely Aurea Lopes Coelho (apoio às coleções)
- Sidnei de Souza (interoperabilidade, interface web)

Informação:

- Arthur Chapman (modelagem, data cleaning, padrões)
- Dora Ann Lange Canhos (web site, relatórios técnicos)
- Ingrid Koch (modelagem, data cleaning)
- Marinez Ferreira de Siqueira (modelagem, relatórios técnicos)
- Rafael Luis Fonseca (apoio às coleções, modelagem)
- Townsend Peterson (modelagem)

Suporte:

- Ana Paula de Souza Albano
- Benedito Aparecido Cruz

Web Design:

- Luísa Donati

Apoio Administrativo:

- Cristina Yoshie Umino
- Silvia Beltrane Lopes

Coleções (verificar)

- **AcariDZSJRP**, Coleção de Ácaros do Departamento de Zoologia e Botânica - IBILCE/UNESP
- **AcariESALQ**, Coleção de Ácaros do Departamento de Entomologia, Fitopatologia e Zoologia - LEF/ESALQ
- **CBMAI**, Coleção Brasileira de Microrganismos de Ambiente e Indústria - CPQBA/UNICAMP
- **DZSJRP**, Coleção de Peixes do Departamento de Zoologia e Botânica - IBILCE/UNESP
- **ESA**, Herbário do Departamento de Ciências Biológicas - LCB/ESALQ
- **IAC**, Herbário do Instituto Agronômico, Campinas - SP
- **IBSBF**, Coleção de Culturas de Fitobactérias do Laboratório de Bacteriologia Vegetal - Instituto Biológico de Campinas

- **LIRP**, Coleção de Peixes do Laboratório de Ictiologia de Ribeirão Preto - FFCLRP/USP
- **SPF**, Herbário do Departamento de Botânica - IB/USP
- **SinBiota**, Sistema de Informação Ambiental - BIOTA/FAPESP