Estudo de Caso
Sistemas de Informação On-line: A experiência do CRIA

Dora Ann Lange Canhos, Sidnei de Souza, Renato de Giovanni, Marinez Ferreira de Siqueira, Alexandre Marino, Rafael Luís Fonseca, Benedito Aparecido Cruz, Vanderlei Perez Canhos.

Centro de Referência em Informação Ambiental

Junho de 2005

Índice

1. Introdução 1
2. Sistemas Centralizados, Distribuídos ou Mistos: Vantagens e Desvantagens 1
   2.1. Sistemas centralizados 1
   2.2. Sistemas distribuídos 2
3. Padrões e Protocolos em Informática para Biodiversidade 3
   3.1. TCS – Taxonomic Concept Transfer Schema 5
   3.2. SDD – Structured Descriptive Data 6
   3.3. DarwinCore 6
   3.4. ABCD – Access to Biological Collection Data 7
   3.5. DiGIR, BioCASe e TAPIR - protocolos para troca de dados 7
4. Exemplo de Sistemas Centralizados no CRIA: SinBiota e SICol 8
   4.1. SinBiota 8
   4.2. SICol 9
5. Exemplo de Sistemas Distribuídos: a Rede *species*Link 11
6. Ferramentas 15
   6.1. MapCRIA 16
   6.2. Data cleaning 17
      a. Erros de Grafia 17
      b. Erros de Coordenadas e de Localidades 19
      c. Geo-referenciamento automático 21
   6.3. *Manager*: Sistema de gerenciamento das coleções participantes 22
      a. Monitor 22
      b. Estatísticas 23
      c. Perfil da Coleção 24
   6.4. OpenModeller: Desenvolvimento de um Ambiente Computacional para Modelagem 28
7. Infra-estrutura 30
   7.1. Hardware 30
   7.2. Software 31
8. Sustentabilidade 31
9. Referências 32

1. Introdução

O Centro de Referência em Informação Ambiental (CRIA), é uma sociedade civil, sem fins lucrativos, que tem como meta e estratégia a disseminação de informação, como ferramenta na organização da comunidade científica e tecnológica do país. Atua especificamente na área de informação biológica, de interesse industrial e ambiental, e pretende, através de sua atuação, contribuir diretamente para a conservação e utilização racional da biodiversidade no Brasil.

A equipe do CRIA trabalha com sistemas de informação on-line desde 1985 quando tornou disponível ao público o Catálogo Nacional de Linhagens através da rede implementada pelo Cirandão, um projeto pioneiro criado pela Embratel, precursor da Internet no Brasil. Essa equipe participou ainda da discussão do Clearing-House Mechanism da Convenção sobre a Diversidade Biológica (CDB) e foi responsável pelo desenvolvimento da Rede Brasileira de Informação em Biodiversidade, a BINbr enquanto fazia parte da Base de Dados Tropical.

Esse conhecimento deu à equipe os subsídios necessários para se responsabilizar pelo desenvolvimento e manutenção de 3 sistemas de informação sobre espécies e espécimes, dois dos quais dão suporte ao programa Biota/Fapesp, O Instituto Virtual da Biodiversidade: o SinBiota e a rede de coleções biológicas *species*Link. O terceiro sistema está voltado a coleções de interesse biotecnológico. Trata-se do SICol (Sistema de Informação de Coleções de Interesse Biotecnológico) desenvolvido com recursos do Ministério da Ciência e Tecnologia e suas agências.

Os três sistemas apresentam várias características distintas já que foram criados em momentos diferentes, como soluções para problemas diferentes. O fato de cada um deles ter sido concebido como um sistema centralizado, distribuído ou misto é o objeto de análise desse documento.

2. Sistemas Centralizados, Distribuídos ou Mistos: Vantagens e Desvantagens

Não se pode dizer *a priori* qual a melhor arquitetura a ser adotada na definição de um sistema de informação. A escolha depende de uma série de fatores como infra-estrutura disponível (*hardware* e *software*), capacitação técnica (*humanware*), conectividade, recursos disponíveis e a "sociologia" da comunidade alvo. Um aspecto essencial é, seja qual for a arquitetura escolhida, o autor ou provedor precisa ter total autonomia e domínio sobre seus dados. Ao autor cabem os créditos e a responsabilidade pela qualidade e veracidade dos dados. Ao gestor do sistema de informação (*custodian*) cabe a tarefa de garantir a integridade dos dados, respeitar eventuais restrições por parte do autor, manter o sistema no ar com *backup* e controles de segurança de rede.

2.1. Sistemas centralizados

Um sistema centralizado (figura 1) caracteriza-se por apresentar dados armazenados em um servidor central. Os autores ou provedores da informação enviam os seus dados ao servidor central seguindo um formato pré-estabelecido. Os dados providos pelos vários participantes da rede, quando armazenados no servidor central, passam a ter a mesma estrutura e formato, constituindo uma base de dados única e homogênea.

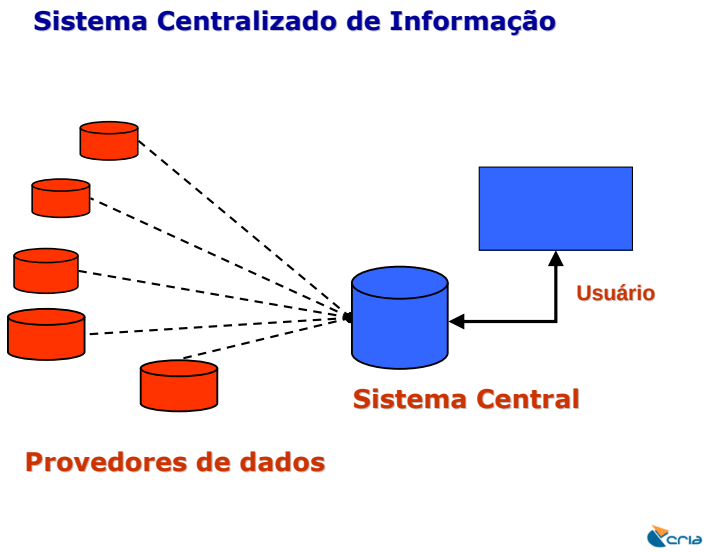


Figura 1. Esquema de um sistema centralizado de informação (CRIA, 2005)

Um sistema centralizado via-de-regra não exige do provedor grande capacitação técnica em informática nem uma infra-estrutura complexa. O fato de estar em um único ambiente, torna o sistema facilmente controlável, o que é uma vantagem do ponto de vista do seu desenvolvimento e manutenção. Mas, uma das principais vantagens é o desempenho. De modo geral, o tempo de resposta de buscas em bases de dados centralizadas é muito menor se comparado ao de bases distribuídas. O trabalho de otimização das rotinas de busca é muito mais fácil de ser tratado, já que depende apenas de fatores internos ao sistema e do tipo de banco de dados utilizado, e não de fatores externos como a performance da rede.

A grande desvantagem é a atualização dos dados. É muito difícil manter uma relação dinâmica entre o provedor da informação e o gestor do sistema, mesmo quando cada provedor é responsável pelos seus dados e não há qualquer interferência ou manipulação dos dados por parte do gestor. O rompimento dessa interação pode inclusive provocar um distanciamento entre o usuário e o provedor de dados já que, com o tempo, os dados podem ficar desatualizados ou até podem não mais responder às perguntas ou atender às necessidades do usuário. A autoria dos dados também pode ser menos evidente em sistemas centralizados. Isso pode trazer algum descontentamento e servir de desestímulo para a manutenção da parceria.

2.2. Sistemas distribuídos

Sistemas distribuídos (figura 2) caracterizam-se por dados armazenados e gerenciados nos servidores dos próprios provedores da informação. Cabe a um portal receber as consultas dos usuários, distribuí-las aos provedores, e depois integrar e devolver os resultados aos usuários.

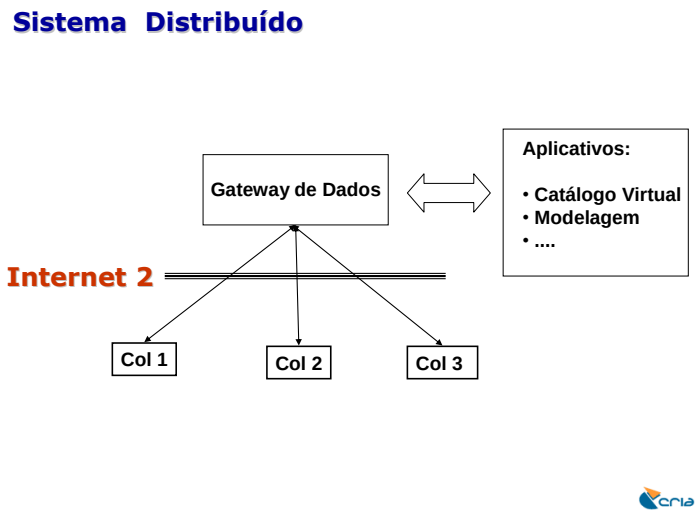


Figura 2. Esquema de um sistema distribuído

Uma grande vantagem de sistemas distribuídos é o fato dos dados estarem sempre atualizados, quase sempre em tempo real, já que o sistema faz acesso ao dado diretamente na sua fonte primária. A autoria e a responsabilidade pelos dados também é evidenciada. Outro aspecto interessante é a co-responsabilidade na manutenção do sistema on-line e a necessidade do estabelecimento de parcerias plenas entre o provedor de dados e o gestor do sistema de informação.

As desvantagens mais importantes incluem a complexidade do sistema, que passa a depender de uma série de fatores externos. O provedor em um sistema distribuído precisa ter uma boa infra-estrutura computacional capacitação em informática, e uma conectividade Internet rápida e estável.

Para minimizar o efeito das variáveis externas, a opção até pouco tempo atrás era a padronização de *hardware* e *software*. Hoje, com o grande avanço no desenvolvimento de padrões e protocolos para a integração de sistemas heterogêneos, é possível integrar sistemas operacionais variados utilizando diferentes software.

3. Padrões e Protocolos em Informática para Biodiversidade

A adoção de padrões e protocolos para a troca de dados e informações sobre biodiversidade é fundamental para a criação de sistemas interoperáveis de informação. De uma maneira geral, podemos definir um padrão como sendo algo definido ou em comum acordo ou por autoridade específica para servir como modelo ou regra para determinado fim. Tem-se também como padrão algo que o consenso geral estabeleceu como modelo. A *World Wide Web Consortium* (W3C)[1], por exemplo, é uma iniciativa que estuda padrões para Web com o objetivo de garantir que as tecnologias fundamentais sejam compatíveis entre si. A idéia é permitir que qualquer hardware ou software utilizado para acessar a Web possam trabalhar em conjunto. A W3C faz referência à sua meta como sendo “interoperabilidade Web”. Através da publicação aberta (não proprietária) de padrões para linguagens e protocolos, estimulando a sua adoção e uso, a W3C busca evitar uma fragmentação que poderia comprometer a Web.

[1] http://www.w3.org/

Um protocolo de comunicação pode ser definido como a descrição formal das regras e formatos de mensagem que dois sistemas devem obedecer para que possam se comunicar e interagir. Talvez os exemplos mais importantes e conhecidos sejam TCP/IP (Transmission Control Protocol / Internet Protocol), SMTP (Simple Mail Transfer Protocol), POP (Post Office Protocol) e IMAP (Internet Message Access Protocol). Esse conjunto de protocolos representa a base de toda a transmissão de dados na Internet desde a troca de emails e a transferência de arquivos até a transmissão de dados em redes distribuídas. Outros padrões de linguagens que também merecem destaque no cenário da Web são HTML (Hyper Text Markup Language) e XML (eXtensible Markup Language) os quais definem as regras de formatação da maioria dos documentos transmitidos através da Internet.

No período que antecedeu a Web, os usuários precisavam de um determinado grau de competência e conhecimento em computação para acessar diferentes sistemas de dados. Graças ao desenvolvimento de padrões e protocolos, hoje, através de apenas um software (web browser) os usuários podem acessar praticamente qualquer sistema de informação disponível na Internet, facilitando sobremaneira o acesso a sistemas verdadeiramente complexos de disseminação de dados e informações. A complexidade passa a estar quase que exclusivamente no desenvolvimento dos sistemas e não no acesso aos dados e informações.

No caso de sistemas de informação para biodiversidade, e principalmente sob o ponto de vista da escala global com inúmeros provedores distribuídos ao redor do mundo servindo uma quantidade muito grande de dados heterogêneos e inter-relacionados, fica evidente que padrões e protocolos precisam ser desenvolvidos e adotados por toda a comunidade. Num futuro muito próximo podemos imaginar uma série de redes temáticas interligadas servindo dados para diferentes públicos e fornecendo ferramentas que beneficiam diretamente os provedores originais (figura 3).

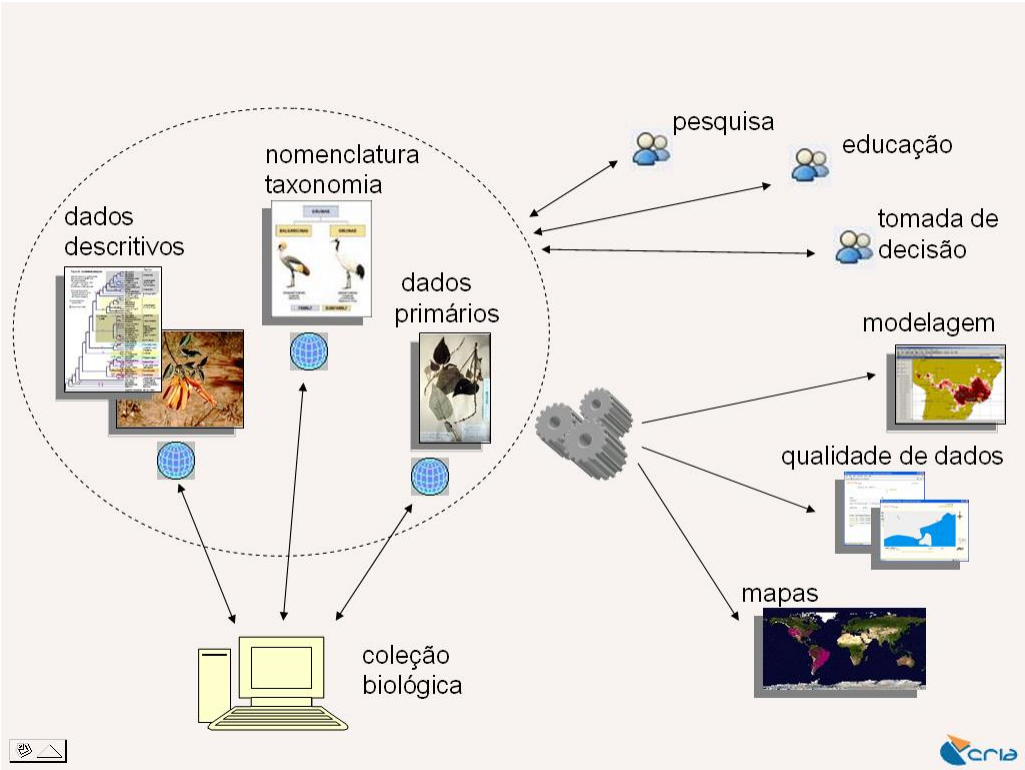


Figura 3. Cenário envolvendo o uso de padrões e protocolos

Um grupo que merece destaque no desenvolvimento de padrões e protocolos para dados sobre espécies e espécimes é o TDWG (*International Working Group on Taxonomic Databases*)[2]. O TDWG tem como missão promover um fórum internacional para projetos sobre dados biológicos, desenvolver e promover o uso de padrões, e facilitar a troca de dados.

São vários grupos de trabalho em atividade que estão buscando estabelecer padrões para:

- Dados de Coleções Biológicas
- Botânica Econômica
- Geografia
- Metadados de Coleções Biológicas
- Dados de imagens e observações
- Padrões para dados espaciais
- Estrutura de dados descritivos de espécies
- Nomes taxonômicos

Alguns padrões que merecem destaque para dados de coleções biológicas são:

- TCS – Taxonomic Concept Transfer Schema
- SDD – Structured Descriptive Data
- DarwinCore
- ABCD – Access to Biological Collection Data
- DiGIR, BioCASe e TAPIR

3.1. TCS – Taxonomic Concept Transfer Schema[3]

O TCS pretende servir como padrão para troca de dados taxonômicos considerando as abordagens de diferentes grupos de usuários (taxonomistas, nomencladores, ecólogos, etc). Utiliza XML (definido através de XML-Schema), e está neste momento aguardando a homologação pelo TDWG. O TCS está centrado na idéia de "conceito taxonômico", ou seja, a classificação de um grupo de organismos por uma pessoa num determinado momento. Cada conceito taxonômico envolve um nome e uma definição. Neste caso, os conceitos taxonômicos foram classificados em: definição original, revisão, conceito incompleto, agregado de conceitos e conceito nomenclatural (quando há uma referência implícita a todos os conceitos que já usaram um mesmo nome). Documentos no padrão TCS poderão portanto conter conceitos, nomes e as relações taxonômicas e nomenclaturais entre eles. Já existe um protótipo em desenvolvimento cujo objetivo é servir como repositório de conceitos taxonômicos (TOS, Taxonomic Object Service[4]), onde pesquisadores poderão realizar consultas e registrar novos dados. Num futuro próximo, cada conceito taxonômico poderá vir a ter um identificador global único que deverá ser capaz de substituir a utilização de nomes científicos em protocolos de troca de dados.

[2] http://www.tdwg.org/

[3] http://www.soc.napier.ac.uk/tdwg/index.php

[4] http://seek.ecoinformatics.org/Wiki.jsp?page=SeekTaxonTools

3.2. SDD – Structured Descriptive Data[5]

O SDD pretende ser um padrão para armazenamento e troca de dados descritivos de organismos (taxa e espécimes). Também utiliza o XML, definido através do XML-Schema. Documentos do tipo SDD poderão armazenar os seguintes dados:

- Metadados sobre o documento
- Terminologia de dados descritivos (em múltiplas línguas e tendo em vista múltiplos públicos-alvo)
- Possibilidade de descrever categorias de organismos (Taxa) ou organismos específicos (espécimes / linhagens)
- Descrições em linguagem natural com possibilidade de marcar texto
- Descrições codificadas
- Chaves de identificação
- Recursos adicionais (glossário, imagens, notas, referências, etc)

Esse padrão também está aguardando homologação pelo TDWG, sendo que já existe um protótipo para editar documentos SDD.

A existência de um padrão para dados descritivos permitindo a integração de dados de diferentes fontes ao redor do mundo deverá facilitar enormemente o processo de identificação e mesmo de descrição de novas espécies.

3.3. DarwinCore[6]

A idéia do DarwinCore foi reunir os elementos (campos) comuns a todos os grupos taxonômicos para padronizar a integração de dados primários de coleções biológicas. Também utiliza XML (definido através de XML-Schema) e aceita extensões. A versão atual do modelo de dados está sendo utilizada pela maioria das redes, inclusive pela rede *species*Link, pelo GBIF[7], pela rede de Mamíferos *Manis*[8], e pela rede OBIS (Ocean Biogeographic Information System[9]), entre outras.

Os campos definidos na versão atual do DarwinCore são: InstitutionCode, CollectionCode, CatalogNumber, ScientificName, BasisOfRecord, Kingdom, Phylum, Class, Order, Family, Genus, Species, Subspecies, ScientificNameAuthor, IdentifiedBy, YearIdentified, MonthIdentified, DayIdentified, TypeStatus, ColectorNumber, FieldNumber, Collector, YearCollected, MonthCollected, DayCollected, JulianDay, TimeOfDay, ContinentOcean, Country, StateProvince, County, Locality, Longitude, Latitude, CoordinatePrecision, BoundingBox, MinimumElevation, MaximumElevation, MinimumDepth, MaximumDepth, Sex, Preparationtype, IndividualCount, PreviousCatalogNumber, RelatedCatalogNumber, RelatedCatalogItem, RelationshipType, Notes, DateLastModified.

Uma nova versão[10] está sendo discutida para ser homologada pelo TDWG.

---

[5] http://160.45.63.11/Projects/TDWG-SDD/
[6] http://darwincore.calacademy.org
[7] http://www.gbif.net
[8] http://elib.cs.berkeley.edu/manis/
[9] http://www.iobis.org/
[10] http://darwincore.calacademy.org/

3.4. ABCD – Access to Biological Collection Data[11]

O objetivo do ABCD foi o de estabelecer um padrão para a troca de dados e metadados de registros em coleções biológicas procurando englobar as particularidades de todos os grupos taxonômicos. O objetivo é idêntico ao DarwinCore só que muito mais detalhado, uma vez que possui cerca de 500 elementos, contra os cerca de 50 elementos do DarwinCore. O modelo de dados ABCD contém elementos específicos para os seguintes tipos de coleções:

- Herbários e Jardins Botânicos
- Coleções Zoológicas
- Coleções de Culturas
- Coleções Paleontológicas

Esse modelo está sendo utilizado pela rede de coleções européias: BioCASE[12]. Como os demais padrões, utiliza XML (definido através de XML-Schema) e está aguardando a homologação pelo TDWG.

DarwinCore e ABCD são os modelos de dados para coleções biológicas sendo adotados pelas principais redes na Internet.

3.5. DiGIR[13], BioCASe[14] e TAPIR[15] - protocolos para troca de dados

As atuais redes que servem dados de coleções biológicas, além de um modelo de dados padrão (como DarwinCore ou ABCD) precisam também de um protocolo para transferência dos dados.

A primeira rede de coleções biológicas a desenvolver um sistema distribuído foi a rede Species Analyst com o uso do protocolo Z39.50 no final dos anos 90. ANSI/NISO Z39.50 é um protocolo utilizado para interconectar sistemas abertos. A primeira versão do padrão foi aprovada em 1988 e é utilizado principalmente por bibliotecas e editoras. O protocolo define o padrão de comunicação entre computadores para a recuperação de informação. Uma característica importante é o fato do Z39.50 suportar ambientes cliente-servidor o que permite separar a interface do usuário (do lado do cliente) do servidor de dados e de ser multiplataforma. Para coleções biológicas ele provou ser muito complexo, exigindo adaptações por parte do provedor de dados.

Dentro do escopo do TDWG, a equipe da Universidade de Kansas, responsável pelo desenvolvimento da rede Species Analyst, e pesquisadores da Universidade da Califórnia e da Academia de Ciências da Califórnia começaram a discutir o desenvolvimento de um outro protocolo mais simples, que atendesse a demanda de uma rede distribuída de dados de coleções biológicas. Optaram por desenvolver esse protocolo de forma cooperativa e colaborativa e lançaram o primeiro código no SourceForge, um ambiente para desenvolvimento de software de código aberto. A equipe do CRIA, que estava iniciando os trabalhos de desenvolvimento da rede *species*Link decidiu participar do desenvolvimento colaborativo ao invés de criar um protocolo próprio. Foi dessa iniciativa que nasceu o protocolo DiGIR (Distributed Generic Information Retrieval). Os requisitos e objetivos da proposta original incluíam:

---

[11] http://www.codata.org/taskgroups/TGbiocollection/
[12] http://www.biocase.org/
[13] http://www.digir.net/
[14] http://www.biocase.org/dev/protocol/index.shtml
[15] http://ww3.bgbm.org/tapir

- Utilização de padrões e protocolos abertos: HTTP, XML e UDDI
- Separação clara entre protocolo, software e semântica
- Facilidade na instalação e configuração de provedores de dados
- Desenvolvimento colaborativo (modelo “open source”)
- Produtos disponíveis a todos através de licença pública (GPL - GNU General Public License)

O desenvolvimento dos trabalhos foi financiado pela NSF (National Science Foundation) nos Estados Unidos e pela Fapesp (Fundação de Amparo à Pesquisa do Estado de São Paulo) no Brasil.

Entretanto, a necessidade de viabilizar a troca de dados utilizando esquemas conceituais mais complexos (no caso o ABCD) levou a rede de coleções Européia (BioCASE) a modificar o protocolo DiGIR e criar um outro protocolo conhecido hoje como *BioCASe*. Infelizmente, “derivações” deste tipo dificultam a interoperabilidade entre sistemas e normalmente acarretam duplicidade de esforços.

Em 2004 foi feito um estudo financiado pelo GBIF para desenvolver um novo protocolo que atendesse às necessidades tanto das redes DiGIR como da rede BioCASE (Döring & Giovanni, 2004). Esse novo protocolo foi denominado TAPIR (TDWG Access Protocol for Information Retrieval) e deve ser lançado ainda em 2005. Espera-se que as redes atuais gradativamente passem a usar o novo protocolo.

4. Exemplo de Sistemas Centralizados no CRIA: SinBiota e SICol

4.1. SinBiota[16]

O SinBiota foi concebido em 1999 com a função de ser um repositório dos dados das coletas realizadas no âmbito do programa Biota/Fapesp. Foi desenvolvido um banco de dados centralizado com alimentação remota, onde o pesquisador pudesse depositar os seus dados, a ficha de coleta e a lista de espécies associadas à coleta. O sistema foi desenvolvido através de uma parceria entre o CRIA, a Unicamp (Institutos de Computação e de Geociências e a Faculdade de Engenharia Agrícola) e o Instituto Florestal, responsável pela digitalização da base cartográfica do estado de São Paulo (escala de 1:50.000).

Para este caso, uma arquitetura centralizada é uma boa opção, uma vez que pesquisadores individuais ou até mesmo grupos de pesquisa via-de-regra não têm estrutura ou interesse em manter um sistema de informação de acesso permanente disponível na Internet. No SinBiota somente pesquisadores cadastrados, associados a algum projeto do programa, podem inserir, corrigir ou até remover seus dados do sistema central. O acesso é protegido por senhas controladas pelos coordenadores de cada projeto. Para a entrada de dados foi elaborada uma ficha padrão de coleta com campos obrigatórios e opcionais, usando vocabulário controlado[17]. Foi também desenvolvida uma estrutura de banco de dados que pudesse integrar os dados de todos os grupos taxonômicos[18]. A figura 4 apresenta um diagrama da arquitetura do sistema.

[16] http://sinbiota.cria.org.br
[17] http://sinbiota.cria.org.br/info/fichapadrao
[18] http://sinbiota.cria.org.br/info/estruturabd

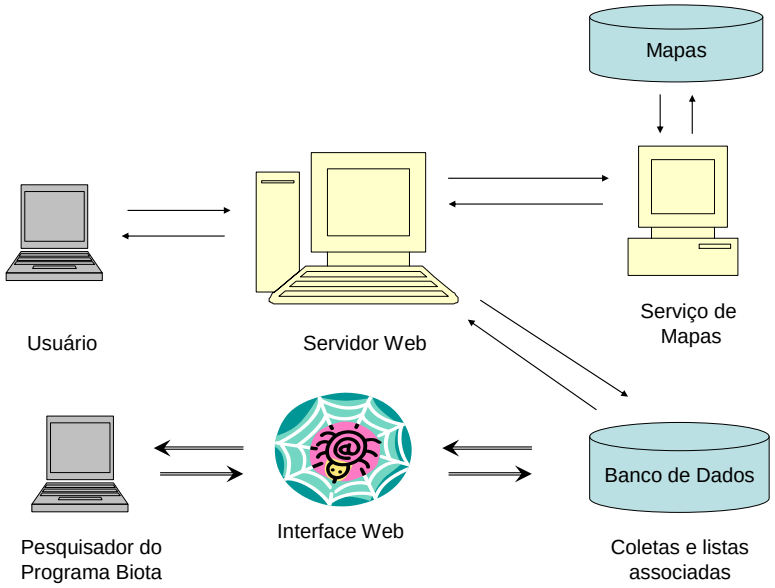


Figura 4. Diagrama da Arquitetura do SinBiota

O sistema está em um servidor Intel/Linux o banco de dados é o PostgreSQL e os bancos de dados secundários estão em XML.

Dados do dia 03 de junho de 2005[19] indicam um total de 7.742 coletas registradas no sistema com cerca de 60 mil espécies associadas a essas coletas. São 60 projetos cadastrados, sendo que nem todos realizam coletas. As estatísticas indicam que o sistema possui 180 usuários responsáveis pela inserção de dados.

4.2. SICol[20]

O segundo sistema centralizado mantido pelo CRIA é o SICol (Sistema de Informação de Coleções de Interesse Biotecnológico), produto de um projeto do Programa de Biotecnologia e Recursos Genéticos do Ministério da Ciência e Tecnologia. Enquanto no SinBiota os dados são enviados por pesquisadores individualmente, no SICol, são enviados em grandes blocos, já organizados e mantidos pelas coleções participantes, através de arquivos pré-formatados. O SICol adotou o padrão CABRI (*Common Access to Biological Resources and Information*)[21], com pequenas modificações, como o modelo de dados. Cada um dos provedores deve produzir e formatar uma planilha de dados de acordo com o modelo definido pelo SICol antes de alimentar o sistema central. Para o envio dos dados, foi criada uma página web através da qual, mediante a utilização de senhas de acesso, as coleções podem periodicamente submeter (enviar) seus dados atualizados. A figura 5 apresenta o esquema adotado pela rede SICol.

[19] http://sinbiota.cria.org.br/info/estatistica
[20] http://sicol.cria.org.br
[21] http://www.cabri.org

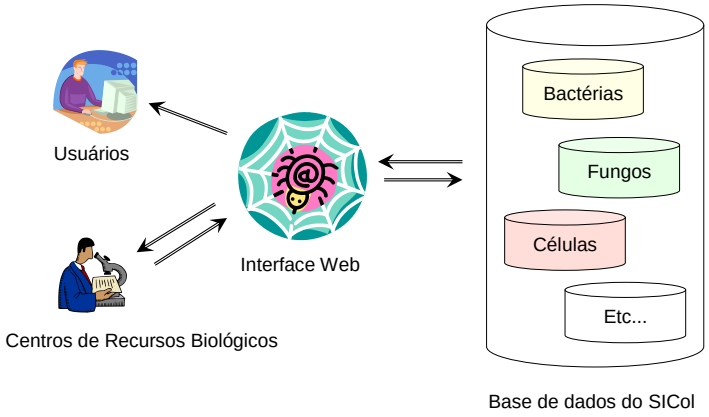


Figura 5. Esquema do SICol

A opção por um sistema centralizado foi feita após a realização de um diagnóstico das coleções quanto à sua infra-estrutura física, existência ou não de pessoal capacitado em informática e conectividade. A absoluta maioria não dispunha nem de infra-estrutura de informática (hardware, software ou "*humanware*"), nem de boa conectividade à Internet. A figura 6 a seguir procura mostrar o sistema implementado.

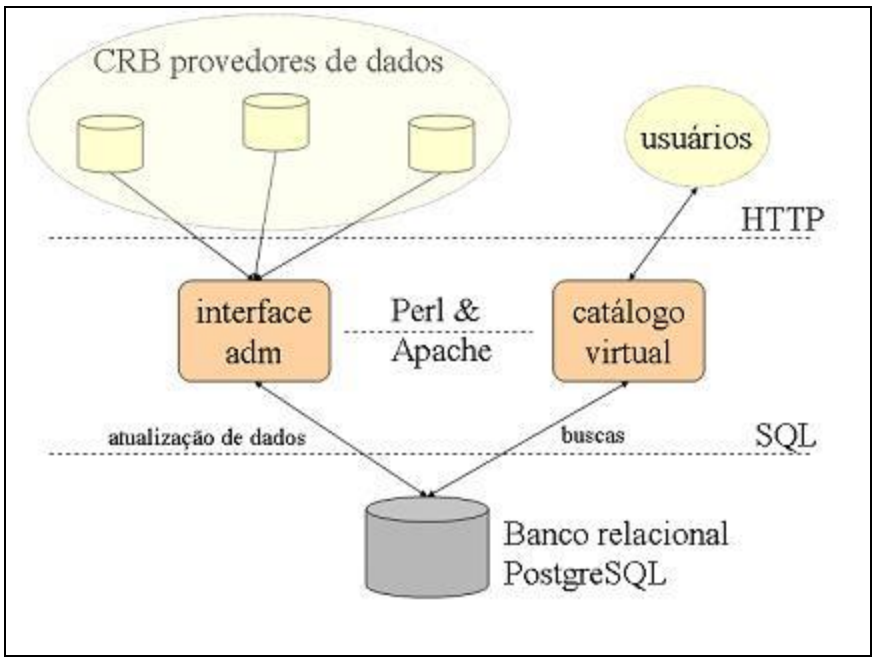


Figura 6. Sistema SICol

O servidor web utilizado pelo SICol é o Apache[22], a implementação do banco de dados foi feita em PostgreSQL[23], e os scripts das páginas foram todos desenvolvidos em linguagem Perl[24].

[22] http://www.apache.org/
[23] http://www.postgresql.org/

Todos são software livre amplamente utilizados e reconhecidos pela comunidade de desenvolvedores.

O catálogo virtual do SICol tem 9 coleções participantes e está disponível on-line. No dia 03 de junho de 2005 disponibilizava 8598 registros. O sistema desenvolvido requer pouco conhecimento por parte da coleção para enviar ou alterar os seus dados. Para alimentar o sistema a coleção precisa exportar seus dados para uma planilha, acessar o sistema usando sua senha e enviar a planilha. No entanto, foi constatado que apesar da simplicidade do processo, foram poucas as coleções que atualizaram seus dados.

5. Exemplo de Sistemas Distribuídos: a Rede *species*Link

A rede *species*Link[25] é um exemplo de um sistema distribuído de dados. O projeto teve por objetivo integrar os acervos de coleções científicas do Estado de São Paulo com os dados armazenados no SinBiota e na rede Species Analyst[26].

O desafio foi integrar os dados das coleções biológicas do Estado de São Paulo interferindo o mínimo possível na sua rotina, adaptando-se aos software já adotados para o gerenciamento dos acervos. (figura 7).

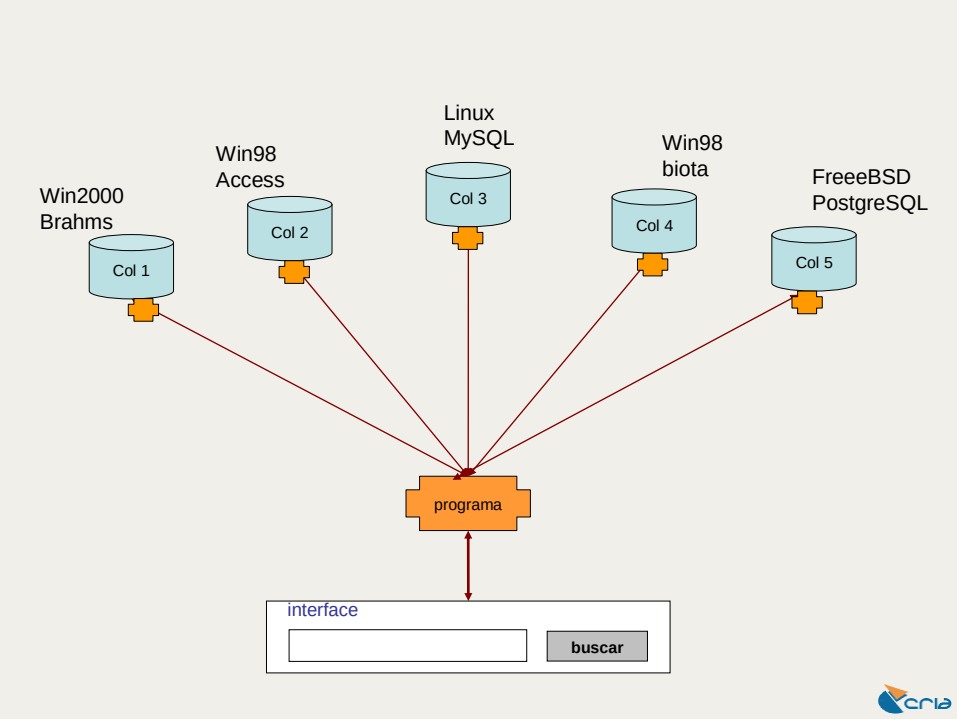


Figura 7. Desafio no desenvolvimento da rede *species*Link

O primeiro passo foi trabalhar com a infra-estrutura de dados, base para todo o desenvolvimento do projeto. Embora grande parte das coleções científicas paulistas tenha se modernizado nos últimos anos devido a incentivos, em especial providos pela Fapesp, a situação entre elas é bastante heterogênea. Foram selecionadas coleções totalmente

[24] http://www.perl.org/
[25] http://splink.cria.org.br/
[26] http://speciesanalyst.net/

informatizadas, parcialmente informatizadas e outras em processo de escolha do *software* a ser utilizado.

Como objeto de pesquisa foi 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 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

As 40 coleções que hoje integram a rede estão utilizando nove software distintos.

Com relação ao protocolo para acesso a dados distribuídos e heterogêneos, o CRIA colaborou no desenvolvimento do protocolo DiGIR, *Distributed Generic Information Retrieval,* um protocolo cliente/servidor já mencionado em seções anteriores, que foi projetado para recuperar informação de fontes distribuídas de acordo com 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 o protocolo para recuperar dados de outros domínios e não apenas de coleções biológicas.

Assim, o DiGIR pode ser entendido como um protocolo configurável, uma vez que as redes que o utilizam podem escolher e definir esquemas conceituais de dados que desejam utilizar. Porém, com vistas a maximizar a interoperabilidade com outras redes, é necessário não apenas adotar o mesmo protocolo mas também um esquema conceitual comum. Foi com este objetivo que foi criado um esquema conceitual genérico, um modelo de dados para coleções biológicas chamado DarwinCore, também descrito em outra seção desse trabalho.

A arquitetura típica de uma rede DiGIR envolve ao menos três componentes distintos:

- Camada de apresentação: É o software que interage com o usuário oferecendo uma interface amigável para especificação de buscas e exibição dos resultados. A camada de apresentação comunica-se com a camada seguinte.
- Camada de distribuição de mensagens (portal): É o software que recebe requisições da camada de apresentação e as distribui para cada um dos provedores de dados conectados à rede. A comunicação com os provedores é feita através do protocolo DiGIR.
- Provedor: É o software responsável por receber requisições do portal e traduzí-las para a linguagem de busca utilizada pelo banco de dados local. O processo de tradução da busca inclui o mapeamento que o provedor fez com relação a um ou mais esquemas conceituais utilizados pelas redes em que participa.

A idéia original seria conectar as coleções diretamente ao portal através desse protocolo. No entanto, no Estado de São Paulo (área de desenvolvimento do protótipo) a maioria das coleções não possui servidor ou rede Internet de alta velocidade nem equipe técnica capaz de manter um sistema de informação permanentemente no ar. A solução foi desenvolver servidores regionais que espelham os dados existentes nas coleções (figura 8).

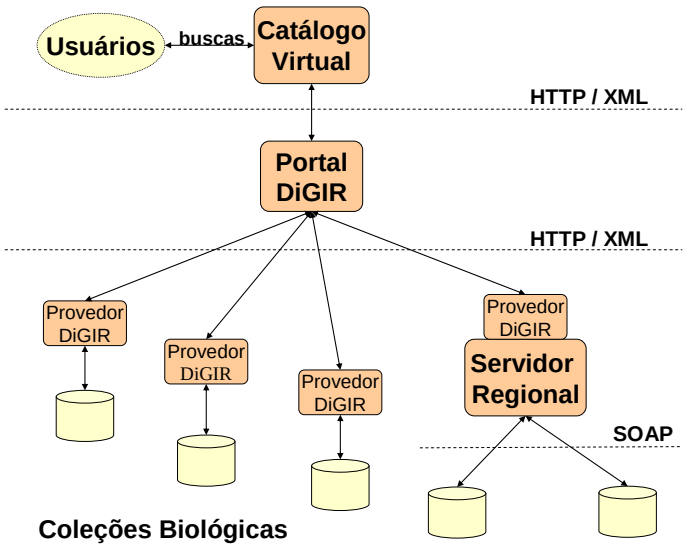


Figura 8. Esquema de um modelo híbrido

Para viabilizar esta arquitetura foram desenvolvidas interfaces, programas capazes de ler os registros e atualizar os bancos de dados nos servidores regionais através de um simples comando de atualização. É possível também desenvolver filtros que dão ao curador a liberdade de omitir dados sensíveis e dessa forma ter total controle sobre o que será ou não disponibilizado on-line. A figura 9 apresenta um esquema da arquitetura adotada pela rede *species*Link.

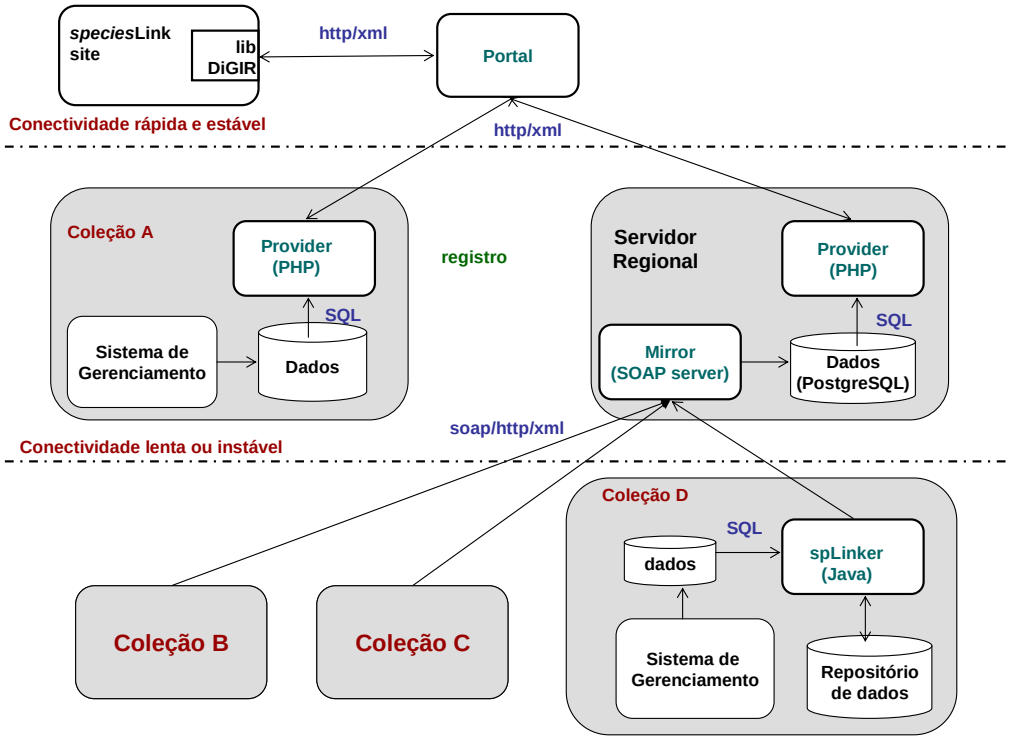


Figura 9. Arquitetura da rede *species*Link

A figura a seguir mostra o diagrama da implementação da arquitetura proposta com as coleções participantes.

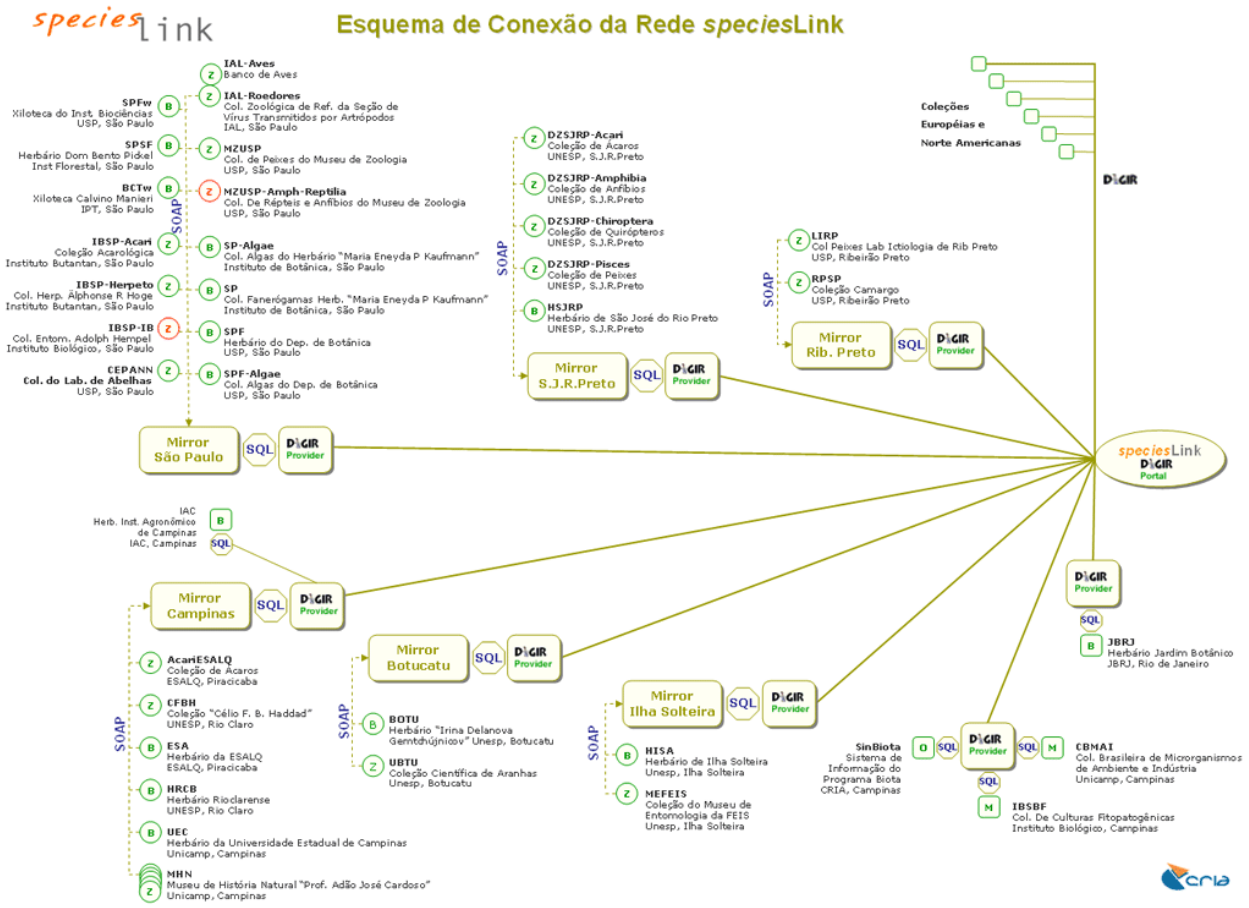


Figura 10. Esquema de conexão da rede *species*Link

No dia 02 de maio de 2005 a rede contava com cerca de 580 mil registros provenientes de 40 coleções brasileiras on-line, todas do Estado de São Paulo com a exceção do Jardim Botânico do Rio de Janeiro. Esse número vem crescendo, comprovando que uma arquitetura que respeita a autonomia das coleções quanto ao controle de seus dados e escolha de seu próprio sistema de gestão está dando resultado. A figura 11 apresenta a entrada e saída de dados da rede mostrando um movimento dinâmico com uma tendência nítida de aumento do acervo disponível.

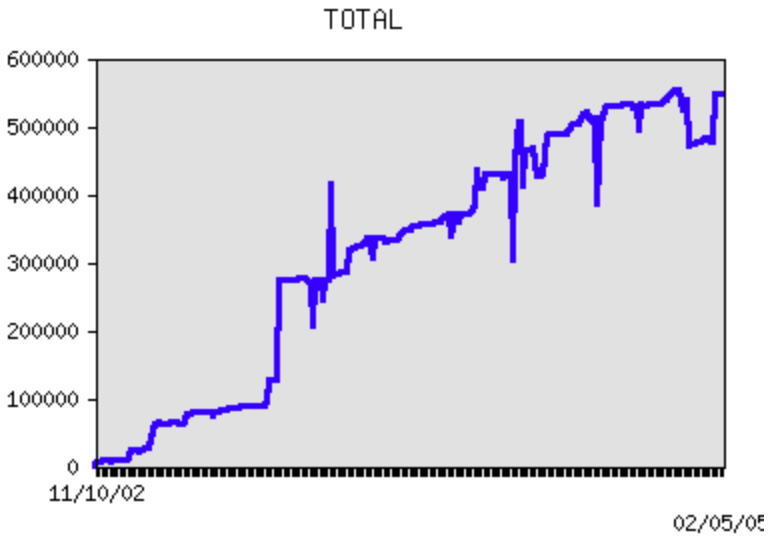


Figura 11. Gráfico da entrada e saída de registros da rede *species*Link obtido dinamicamente[27] no dia 02 de maio de 2005.

Alguns aspectos técnicos da rede speciesLink que merecem destaque são:

- Hardware: a rede usa equipamento Intel (microcomputadores);
- Software: o sistema foi desenvolvido com software livre e de protocolo aberto;
- O protocolo responsável pela interoperabilidade de sistemas é o DiGIR desenvolvido de forma colaborativa no *source forge*;
- O modelo de dados utilizado é o Darwin Core, também objeto de desenvolvimento internacional;
- As coleções têm total autonomia quanto ao sistema operacional e ao software que desejam utilizar localmente;
- As coleções têm total liberdade de inserir ou remover o banco de dados, registros específicos, campos específicos, ou ainda um ou mais campos de um ou mais registros específicos;
- As coleções que não dispõem de acesso rápido à Internet, nem de servidor dedicado na rede, podem participar;
- A rede é de fácil expansão.

6. Ferramentas

Além da disseminação de dados, existem outras vantagens tanto do ponto de vista do provedor como também do usuário de ter dados disponíveis on-line. Na rede speciesLink destacamos as ferramentas como o mapCRIA para a visualização dos dados em mapas, o *data cleaning* para a identificação de registros "suspeitos", o *manager* que monitora os trabalhos da coleção e o openModeller, um ambiente para a modelagem preditiva da distribuição de espécies. O desenvolvimento dessas ferramentas só foi possíveis graças à interação com a comunidade provedora de dados e usuária do sistema.

[27] http://splink.cria.org.br/manager/index?action=stats

6.1. MapCRIA[28]

Desde o início do desenvolvimento dos sistemas de informação para o Programa Biota/Fapesp, foi detectada a necessidade de um aplicativo para a produção dinâmica de mapas na internet. Foram várias versões, passando desde o uso de software proprietário (ArcInfo versão Unix) até a solução atual utilizando MapServer, um pacote de código aberto desenvolvido pela Universidade de Minnesota (UMN) em cooperação com a NASA. O MapServer foi escolhido por ser de código aberto, ter desenvolvimento colaborativo, ser multi-plataforma, e pelo fato de também disponibilizar uma biblioteca que é utilizada como base para o desenvolvimento de aplicações desenhadas especificamente para as necessidades dos sistemas desenvolvidos pelo CRIA, o MapScript.

Foi implementado um serviço web padronizado que faz a interface entre os diferentes aplicativos desenvolvidos e a biblioteca MapScript. Foi também desenvolvida uma aplicação padrão capaz de receber parâmetros de mapas previamente inicializados pelo serviço de mapas que tivesse autonomia para continuar a interação com o usuário provendo funções básicas como *zoom* (*in/out*), ligar e desligar camadas (*layers*), entre outras. Além disso, o aplicativo deveria ser suficientemente leve na rede e ser compatível com os navegadores mais utilizados. A interface foi desenvolvida utilizando apenas DHTML e JavaScript, sem a utilização de Java.

Foram então implementados um serviço web, por nós chamado de ***mapcria web service***, e uma visualizador, batizado de ***mapcria viewer***. A figura a seguir apresenta a idéia de interação básica entre os módulos e as aplicações que os utilizam.

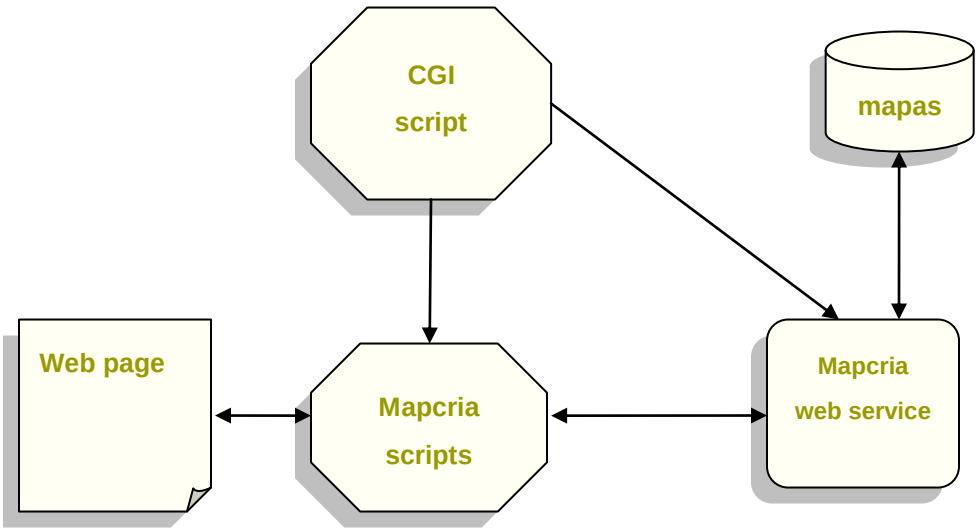


Figura 12. Interação entre os módulos e as diferentes aplicações do sistema web desenvolvido pelo CRIA

O serviço Mapcria para construção e manipulação de mapas, está na sua versão 2.0 e é baseado na biblioteca MapScript C com interface Perl SWIG para MapServer 4.2, utilizando ainda SOAP::Lite. O serviço está disponível em ***mapcria.cria.org.br*** porta ***59000***.

[28] http://www.cria.org.br/mapcria/

O mapcria viewer pode ser entendido com um conjunto de aplicações (CGI Perl scripts) capazes de interagir com o serviço mapcria utilizando as várias funções disponibilizadas pelo serviço através do navegador do usuário, permitindo assim a interatividade entre o usuário e o serviço através de um browser.

O módulo principal é responsável por criar a página web onde o mapa será apresentado ao usuário, assim como gerar o JavaScript específico para aquele caso e ambiente. É importante ressaltar que todo o código necessário para criar a página apresentada ao usuário é gerado dinamicamente para se adaptar às características tanto do mapa quanto do ambiente utilizado pelo usuário (tipo de navegador, versão, sistema operacional, tamanho de tela, etc.). Alguns exemplos de utilização do visualizador serão apresentados com a ferramenta *data cleaning.*

6.2. Data cleaning[29]

Esta ferramenta foi desenvolvida para auxiliar as coleções no processo de verificação e correção de erros, na complementação de dados e na análise do processo de informatização. O sistema indica quais os registros suspeitos, cabendo ao provedor do dado avaliar e eventualmente corrigir registros com erros. A ferramenta não modifica nenhum dado.

a. Erros de Grafia

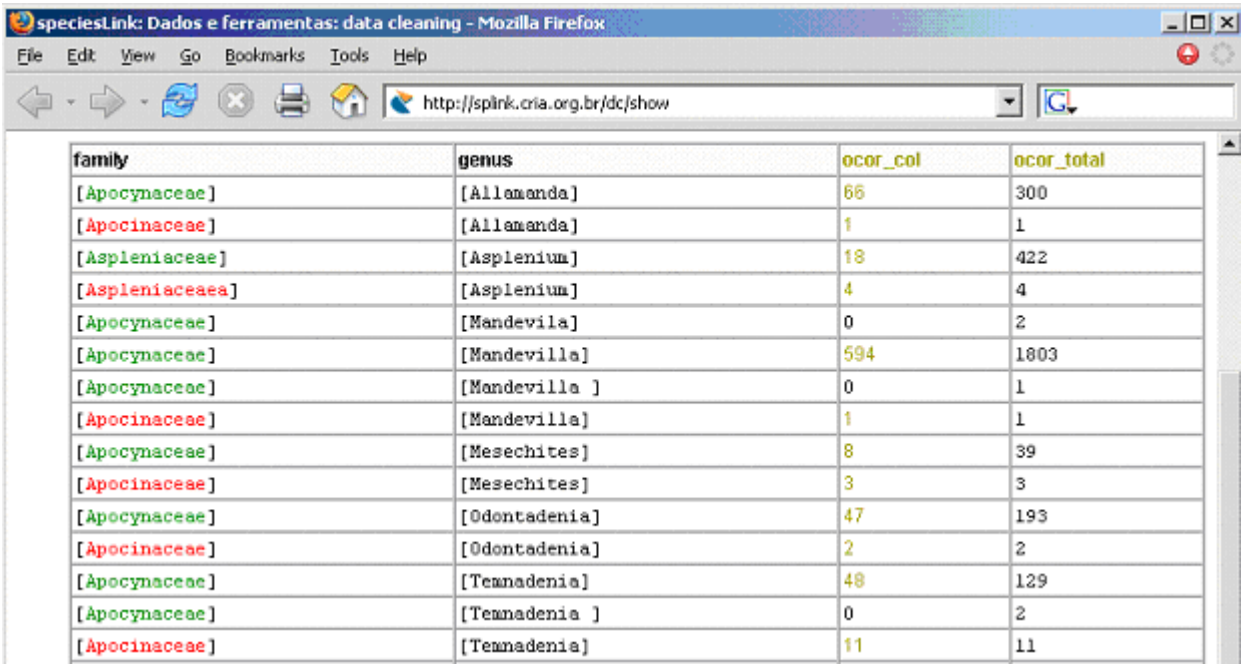
O sistema realiza uma checagem dos campos de família, gênero, espécie e autor, comparando-os e fazendo algumas suposições. Se um registro tem o mesmo nome para família e espécie, por exemplo, o sistema supõe que o gênero deve ser o mesmo. É feita uma busca fonética e quando há uma variação na grafia os registros são apresentados como sendo "suspeitos". Para cada registro suspeito é indicado o número de ocorrências daquele conjunto na coleção e em toda a rede *species*Link. O sistema também indica se o nome consta no Catálogo da Vida do Species 2000[30]. Outras listas de referência poderiam ser utilizadas, principalmente sobre espécies brasileiras, desde que disponibilizadas eletronicamente.

A seguir são apresentados alguns exemplos de registros “suspeitos” para ilustrar o conceito. A tabela 1, por exemplo, registra uma variação de grafia para a família Apocynaceae.

[29] http://splink.cria.org.br/dc/

[30] http://www.sp2000.org

Tabela 1. Exemplos de nomes suspeitos de famílias



family	genus	ocor_col	ocor_total
[Apocynaceae]	[Allamanda]	66	300
[Apocinaceae]	[Allamanda]	1	1
[Aspleniaceae]	[Asplenium]	18	422
[Aspleniaceaea]	[Asplenium]	4	4
[Apocynaceae]	[Mandevila]	0	2
[Apocynaceae]	[Mandevilla]	594	1803
[Apocynaceae]	[Mandevilla]	0	1
[Apocinaceae]	[Mandevilla]	1	1
[Apocynaceae]	[Mesechites]	8	39
[Apocinaceae]	[Mesechites]	3	3
[Apocynaceae]	[Odontadenia]	47	193
[Apocinaceae]	[Odontadenia]	2	2
[Apocynaceae]	[Temnadenia]	48	129
[Apocynaceae]	[Temnadenia]	0	2
[Apocinaceae]	[Temnadenia]	11	11

O sistema indica que o nome (família e gênero) Apoc**y**naceae e Allamanda, por exemplo ocorre 66 vezes no acervo analisado e 300 vezes na rede *species*Link. No entanto, o nome Apoc**i**naceae e Allamanda ocorre apenas 1 vez no acervo analisado e é a única ocorrência em toda a rede *species*Link. Os nomes suspeitos que aparecem em vermelho são nomes que não constam nos dicionários disponíveis no CRIA, já os em verde constam. Portanto, comparando o número de ocorrências na própria coleção e em toda a rede, o sistema procura indicar que o nome escrito com "y" tem maior probabilidade de estar correto. É importante salientar que o sistema não altera os dados, procurando apenas dar elementos para que o responsável pela informação possa decidir se o registro "suspeito" está realmente errado ou não. Se a grafia correta for Apoc**y**naceae o curador pode *clicar* na ocorrência do nome Apoc**i**naceae para identificar o número do registro na coleção:

catalognumber	family	genus	species	subspecies	scientificnameauthor	collectornumber	collector	fieldnumber
67017	Apocinaceae	sp Allamanda	cathartica				Martins, AB	7966
total de registros: 1								

Nota: algumas colunas foram excluídas do registro por uma questão de formatação

Neste caso, ao curador basta acessar o registro 67017 de seu banco de dados e alterar o nome da família de Apocinaceae para Apocynaceae.

O mesmo conceito é aplicado para nomes de gênero e espécies.

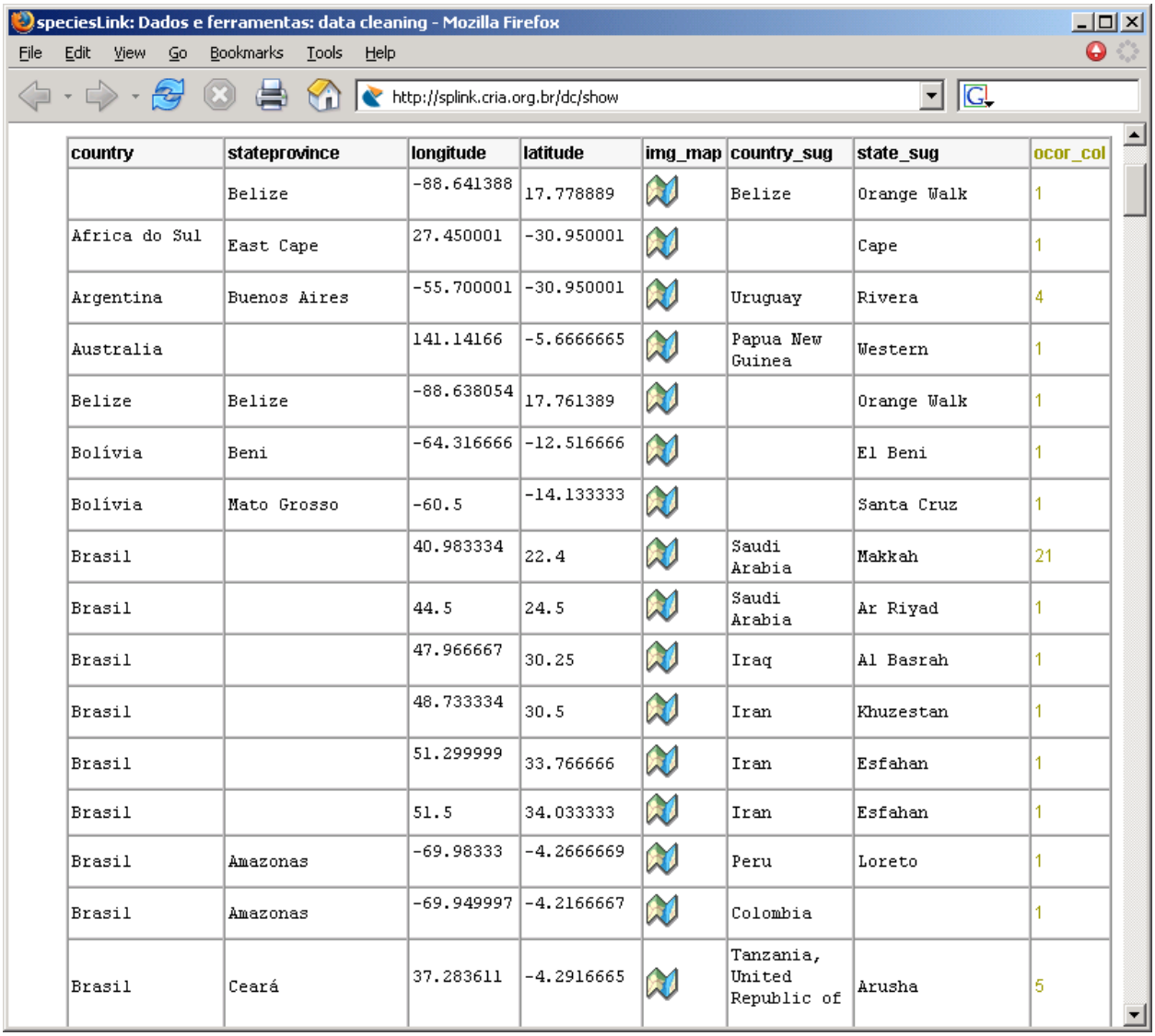
Essa ferramenta mostra a importância da existência de *checklists* de espécies locais com nomes validados por especialistas. É fundamental que as informações de iniciativas como a Flora Fanerogâmica do Estado de São Paulo e outras listas de nomes validados sejam rapidamente disponibilizadas *on-line*.

b. Erros de Coordenadas e de Localidades

O sistema compara a latitude e longitude com o nome de país, estado e município indicados pela coleção, procurando inconsistências. Como fonte de dados esta ferramenta utiliza a base de dados de localidades brasileiras do IBGE. O sistema ainda identifica *outliers* usando técnicas modificadas por Chapman 1999 (Chapman, 1999) para detectar os pontos fora do padrão esperado para os parâmetros latitude, longitude e altitude. São também verificados os registros com coordenadas geográficas fora do limite mundial, com latitude e/ou longitude igual a zero, além de pontos fora do limite do mar territorial brasileiro (quando o campo do país é o Brasil).

A tabela 3 apresenta alguns exemplos de registros “suspeitos” em relação à localização geográfica da informação.

Tabela 2. Lista de registros com provável erro de geo-referenciamento.



country	stateprovince	longitude	latitude	img_map	country_sug	state_sug	ocor_col
	Belize	-88.641388	17.778889		Belize	Orange Walk	1
Africa do Sul	East Cape	27.450001	-30.950001			Cape	1
Argentina	Buenos Aires	-55.700001	-30.950001		Uruguay	Rivera	4
Australia		141.14166	-5.6666665		Papua New Guinea	Western	1
Belize	Belize	-88.638054	17.761389			Orange Walk	1
Bolívia	Beni	-64.316666	-12.516666			El Beni	1
Bolívia	Mato Grosso	-60.5	-14.133333			Santa Cruz	1
Brasil		40.983334	22.4		Saudi Arabia	Makkah	21
Brasil		44.5	24.5		Saudi Arabia	Ar Riyad	1
Brasil		47.966667	30.25		Iraq	Al Basrah	1
Brasil		48.733334	30.5		Iran	Khuzestan	1
Brasil		51.299999	33.766666		Iran	Esfahan	1
Brasil		51.5	34.033333		Iran	Esfahan	1
Brasil	Amazonas	-69.98333	-4.2666669		Peru	Loreto	1
Brasil	Amazonas	-69.949997	-4.2166667		Colombia		1
Brasil	Ceará	37.283611	-4.2916665		Tanzania, United Republic of	Arusha	5

Os dados à esquerda da tabela correspondem aos dados registrados na coleção. O mapa e os dados à direita do ícone “mapa” são gerados pelo sistema. A tabela indica, por exemplo alguns registros que a coleção registra como ocorrendo no Brasil cujos valores de longitude e latitude,

quando analisados pelo sistema, indicam a ocorrência da coleta em países como Arábia Saudita, Iraque e Iran. Esse é um exemplo relativamente comum de omissão do sinal (-) nos valores registrados.

A mesma técnica utilizada para identificar os países suspeitos é utilizada para identificar os municípios suspeitos. Nesse caso somente são checados os registros que ocorrem no Brasil. A figura a seguir mostra uma tabela para municípios suspeitos.

country	state	county	longitude	latitude	img_map	state_sug	county_sug	ocor_col
Brasil	[Mato Grosso do Sul]	[Três Lagoas]	-52	-20			Inocência	1
Brasil	[Mato Grosso do Sul]	[Três Lagoas]	-51	-20		MG	Carneirinho	10

Para o segundo registro a coleção indica que a coleta foi realizado no município de Três Lagoas no estado do Mato Grosso do Sul e o sistema está indicandicando que o ponto cai no município de Carneirinho em Minas Gerais. Clicando no ícone do mapa temos:

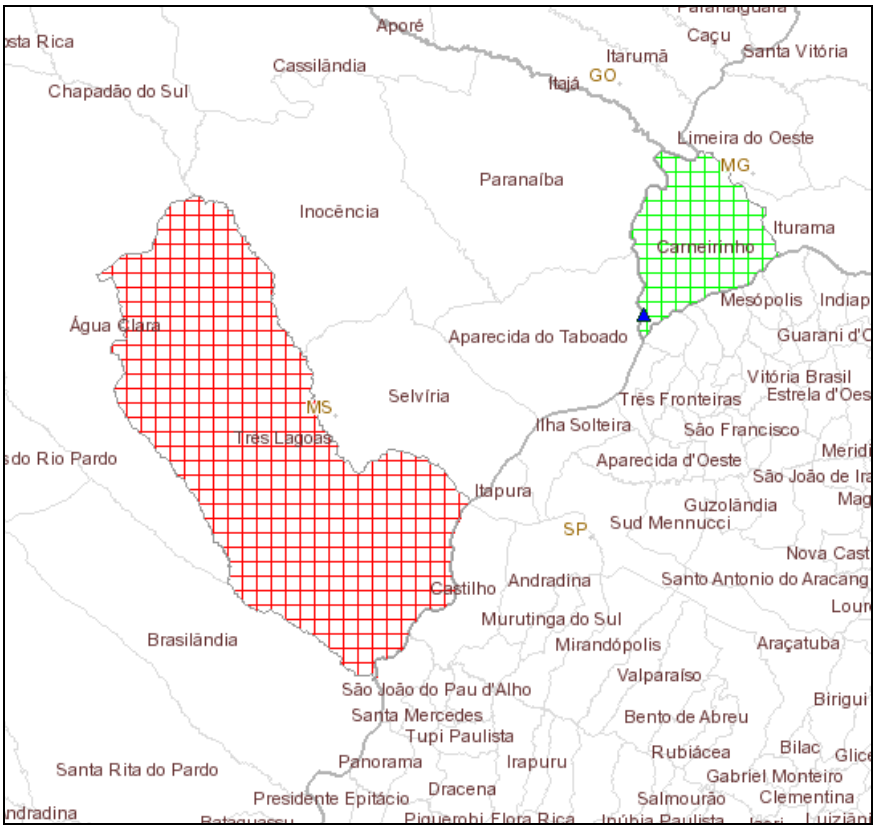


Figura 13. Localização do ponto indicado

O usuário pode ainda adicionar mais informações (*layers*) aos mapas como rodovias e rios para auxiliar na avaliação do curador quanto à localização correta do ponto.

Para a detecção dos *outliers* a ferramenta analisa todos os pontos geo-referenciados na coleção e utiliza técnicas estatísticas para identificar aqueles que estão fora do padrão esperado.

c. Geo-referenciamento automático

O geo-referenciamento automático tem por objetivo sugerir valores de longitude e latitude para registros que possuem dados sobre a localidade, como, por exemplo, o nome do município. A fonte dos dados é a base de dados de localidades brasileiras do IBGE. Como essas coordenadas não são precisas, recomenda-se indicar a precisão no registro de dados. Informar isso é importante para que o usuário possa decidir sobre o uso ou não desse dado. Dependendo do tipo de uso, essa informação pode ser suficientemente precisa. É importante ressaltar que para algumas coleções essa ferramenta chegou a sugerir coordenadas geográficas para mais de 80% dos registros sem coordenadas.

A figura 14 mostra o mapa produzido automaticamente para uma coleção onde os registros sem coordenadas tinham o nome dos municípios onde as coletas foram realizadas.

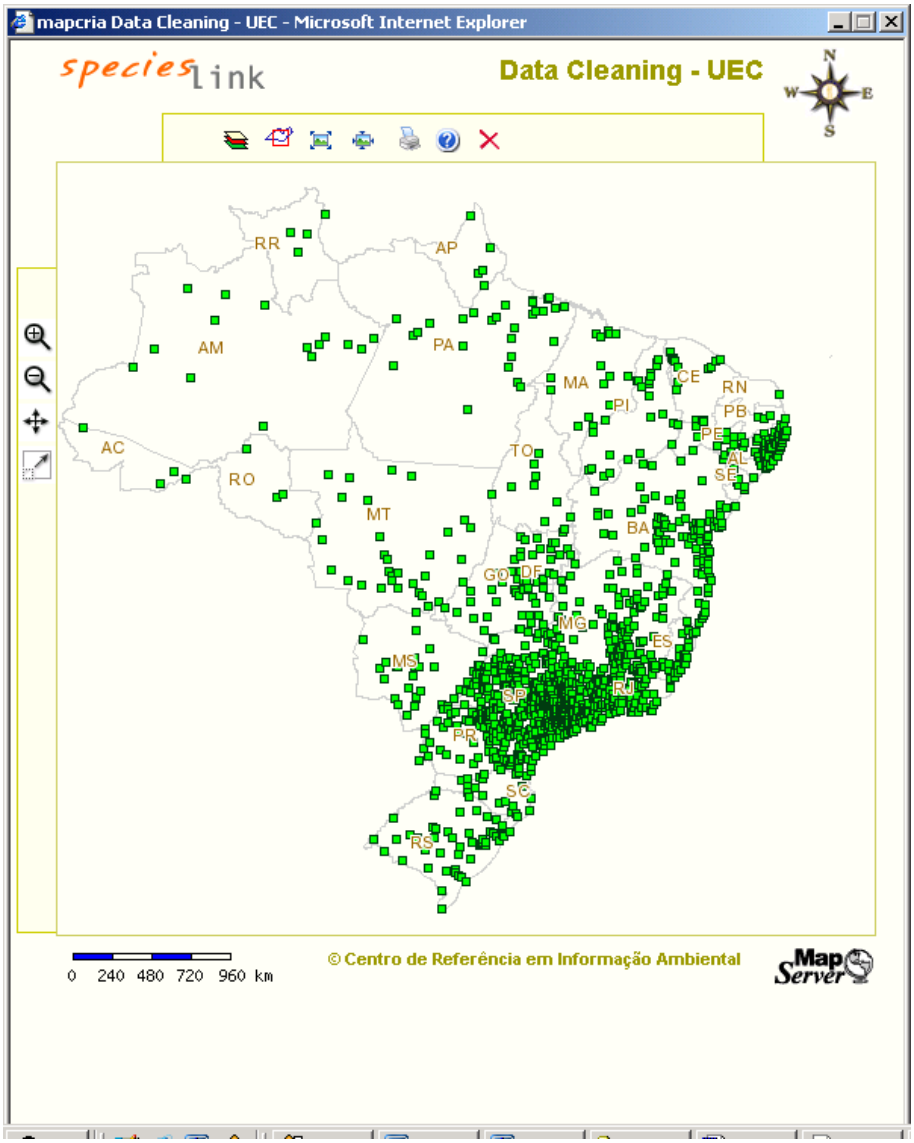


Figura 14. Geo-referenciamento automático de registros sem coordenadas geográficas mas com dados sobre o município da coleta.

6.3. *Manager*: Sistema de gerenciamento das coleções participantes[31]

Com o crescente número de coleções participantes da rede *species*Link, principalmente a partir do início da segunda fase do projeto em outubro de 2003, tornou-se imprescindível desenvolver um sistema que permitisse o gerenciamento do *status* de cada uma das coleções. O sistema foi desenvolvido com a intenção de auxiliar no acompanhamento das atividades do projeto, para uso interno do CRIA e do coordenador do projeto. No entanto alguns módulos mostraram ser de interesse mais amplo, pois permitem visualizar a evolução da rede. Destacamos os módulos "monitor", "estatística", e o "perfil da coleção".

a. Monitor

Esse módulo apresenta ao usuário uma tabela contendo o nome de cada coleção, a sigla, a cidade onde está localizada, o número de registros disponíveis para consulta na rede, o número total de registros no acervo e calcula e apresenta também o percentual de registros disponíveis on-line. Além dessas informações, a disponibilidade de conexão com a coleção é sinalizada. Processos de verificação da conectividade e número de registros são executados a cada quarto de hora e alimentam o banco de dados central.

Tabela 3. Informações disponíveis ao público sobre cada coleção

Coleção	sigla	cidade	registros	total	%
Coleção de Ácaros do Departamento de Entomologia, Fitopatologia e Zoologia	AcariESALQ	Piracicaba	12.392	15.000	83%
Coleção Brasileira de Microrganismos de Ambiente e Indústria	CBMAI	Paulínia	314	688	46%
Coleção de Ácaros	DZSJRP-Acari	São José do Rio Preto	5.753	7.000	82%
Coleção de Peixes	DZSJRP-Pisces	São José do Rio Preto	7.441	7.441	100%
Herbário do Instituto Agronômico de Campinas	IAC	Campinas	36.051	45.000	80%
Coleção de Culturas de Fitobactérias do Instituto Biológico	IBSBF	Campinas	1.624	2.000	81%
Coleção de Peixes do Laboratório de Ictiologia de Ribeirão Preto	LIRP	Ribeirão Preto	4.928	30.000	16%
Coleção de Peixes do Museu de Zoologia da USP	MZUSP	São Paulo	72.706	84.000	87%
Sistema de Informação do Programa Biota/Fapesp	SinBiota	Campinas	57.461	57.461	100%
Coleção de Algas do Herbário do Estado "Maria Eneyda P. Kaufmann Fidalgo"	SP-Algae	São Paulo	13.235	15.000	88%
Coleção de Fanerógamas do Herbário do Estado "Maria Eneyda P. Kaufmann Fidalgo"	SP	São Paulo	13.814	350.000	4%
Herbário do Departamento de Botânica, IB/USP	SPF	São Paulo	18.800	133.500	14%
Coleção de Algas do Departamento de Botânica, IB/USP	SPF-Algae	São Paulo	19.776	19.776	100%
Herbário da Universidade Estadual de Campinas	UEC	Campinas	35.382	134.000	26%
Xiloteca Calvino Mainieri	BCTw	São Paulo	3.359	34.500	10%
Herbário "Irina Delanova Gemtchújnicov"	BOTU	Botucatu	0	0	0
Coleção do Laboratório de Abelhas do IB/USP	CEPANN	São Paulo	26.126	0	0
Coleção "Célio F. B. Haddad" CFBH Rio Claro	CFBH	Rio Claro	2.935	7.000	42%
Coleção de plantas medicinais e aromáticas	CPMA	Campinas	1.882	2.150	88%
Coleção de Anfíbios	DZSJRP-Amphibia	São José do Rio Preto	7.146	7.146	100%
Coleção de Quirópteros	DZSJRP-	São José do	10.678	10.678	100%

[31] http://splink.cria.org.br/manager

Coleção	sigla	cidade	registros	total	%
	Chiroptera	Rio Preto			
Herbário da Escola Superior de Agricultura Luiz de Queiroz	ESA	Piracicaba	45.252	120.000	38%
Herbário de Ilha Solteira	HISA	Ilha Solteira	182	10.235	2%
Herbário Rioclarense	HRCB	Rio Claro	2.593	40.000	6%
Herbário de São José do Rio Preto	HSJRP	São José do Rio Preto	19.380	28.000	69%
Coleção Zoológica de Referência da Seção de Vírus Transmitidos por Artrópodos - Banco de Aves	IAL-aves	São Paulo	5.129	110.000	5%
Coleção Zoológica de Referência da Seção de Vírus Transmitidos por Artrópodos	IAL-roedores	São Paulo	10.851	21.000	52%
Coleção Entomológica "Adolph Hempel" do Instituto Biológico	IBSP-IB	São Paulo	0	275.000	0%
Coleção Acarológica do Instituto Butantan	IBSP-Acari	São Paulo	4.210	9.201	46%
Coleção Herpetológica "Alphonse Richard Hoge"	IBSP-Herpeto	São Paulo	57.397	80.000	72%
Herbário Dimitri Sucre Benjamin	JBRJ	Rio de Janeiro	450.000	14	0%
Coleção do Museu de Entomologia da FEIS/Unesp	MEFEIS	Ilha Solteira	3.519	28.000	13%
Coleção de Anfíbios do Museu de História Natural "Prof. Dr. Adão José Cardoso"	MHN-anfibios	Campinas	16.062	16.062	100%
Coleção de Aves do Museu de História Natural "Prof. Dr. Adão José Cardoso"	MHN-aves	Campinas	2.197	2.197 %	100%
Coleção de Mamíferos do Museu de História Natural "Prof. Dr. Adão José Cardoso"	MHN-mamiferos	Campinas	2.359	2.359 %	100%
Coleção de Peixes do Museu de História Natural "Prof. Dr. Adão José Cardoso"	MHN-peixes	Campinas	7.767	7.767	100%
Coleção de Répteis do Museu de História Natural "Prof. Dr. Adão José Cardoso"	MHN-repteis	Campinas	2.439	2.439	100%
Coleção Camargo	RPSP	Ribeirão Preto	39.991	171.000	23%
Xiloteca do Instituto de Biociências da Universidade de São Paulo	SPFw	São Paulo	908	4.000	23%
Herbário Dom Bento Pickel	SPSF	São Paulo	12.735	34.000	37%
Coleção Científica de Aranhas (Araneae) do Depto. de Zoologia da Unesp, Campus Botucatu	UBTU	Botucatu	2.625	3.500	75%
Totais			587.400	2.377.100	25%

A tabela 4 indica que no dia 24 de maio de 2005 às 12:48 a rede *species*Link tinha 587.400 registros on-line, representando 25% do total de registros das coleções participantes.

b. Estatísticas

Esse módulo apresenta um gráfico geral sobre a evolução do número de registros disponíveis na rede desde 11 de outubro de 2002 (figura 15). Apresenta também gráficos individuais com a entrada e saída de dados de cada coleção. Os gráficos são gerados dinamicamente de acordo com os dados coletados a cada 15 minutos pelos processos de verificação.

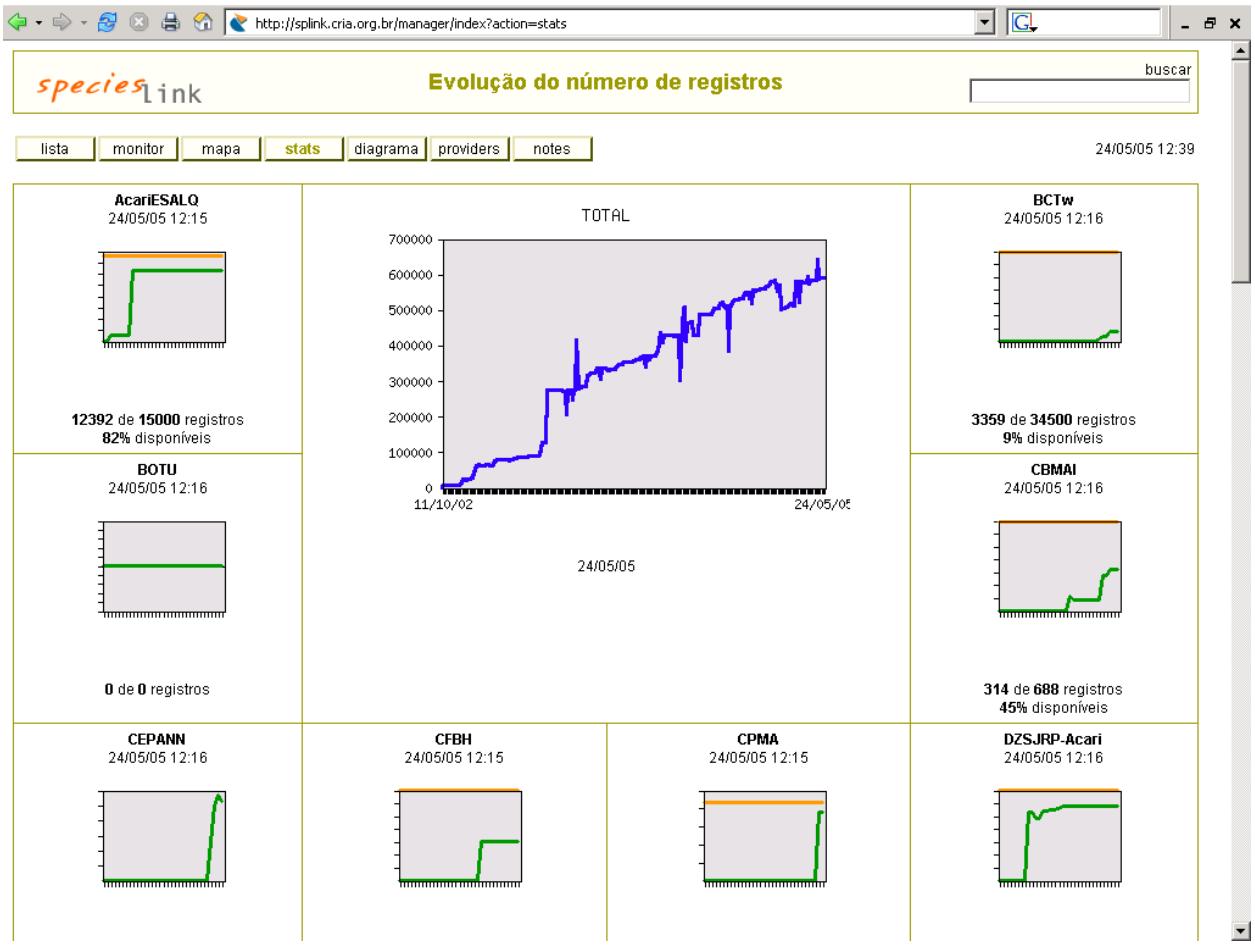


Figura 15. Monitoramento da entrada e saída de dados da rede *species*Link

As estatísticas mostram o nível de atividade de cada coleção com relação à entrada e saída de dados. O gráfico geral mostra que a rede é dinâmica e que as coleções estão tendo plena liberdade na gestão de seus dados na rede.

c. Perfil da Coleção

Cada coleção tem o seu perfil desenhado na rede. São apresentados na figura 16 os gráficos de entrada e saída de dados, um mapa da distribuição de seus registros on-line, além de seus dados cadastrais, pessoas de contato, software utilizado, número total de registros e descrição do acervo.

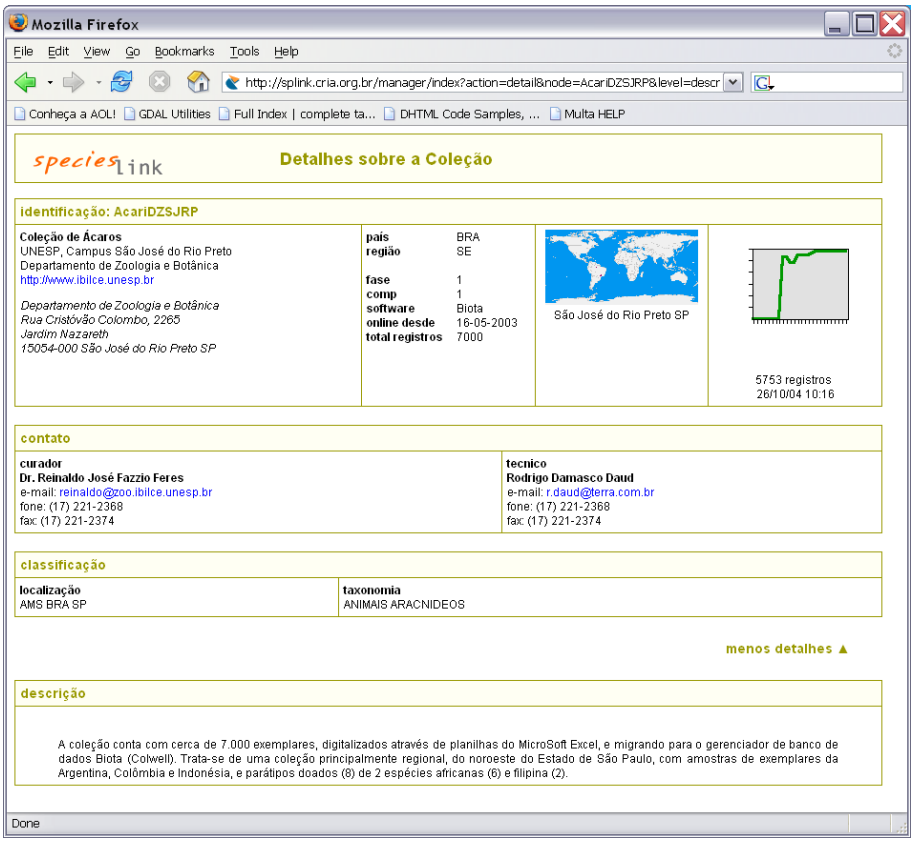
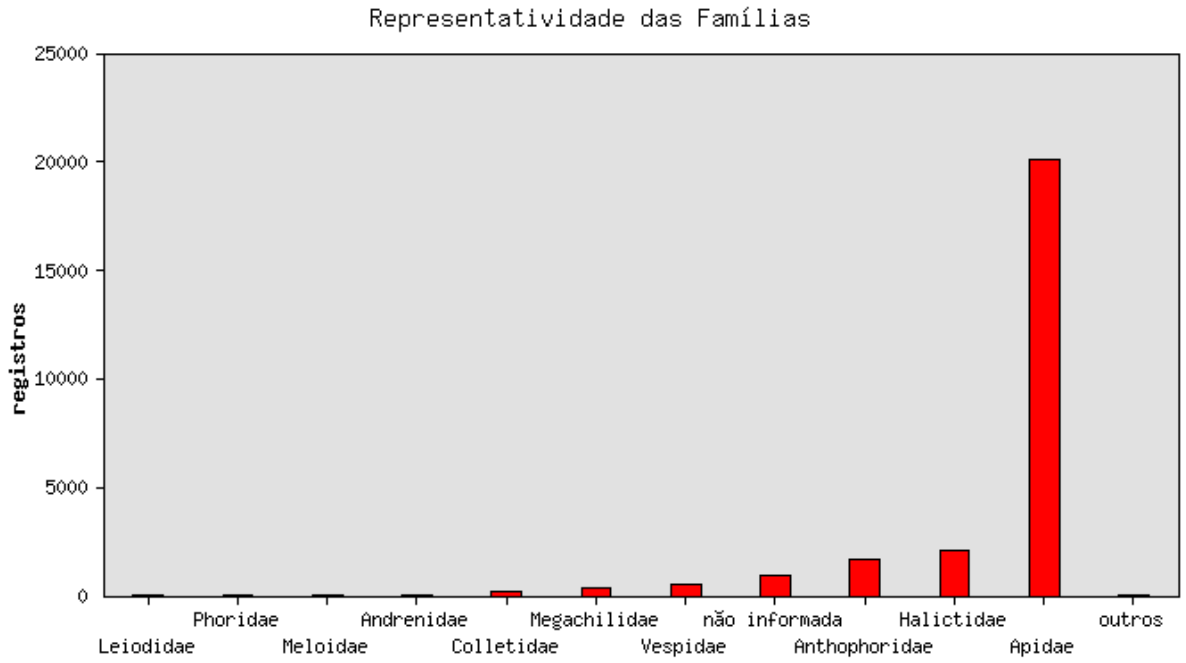
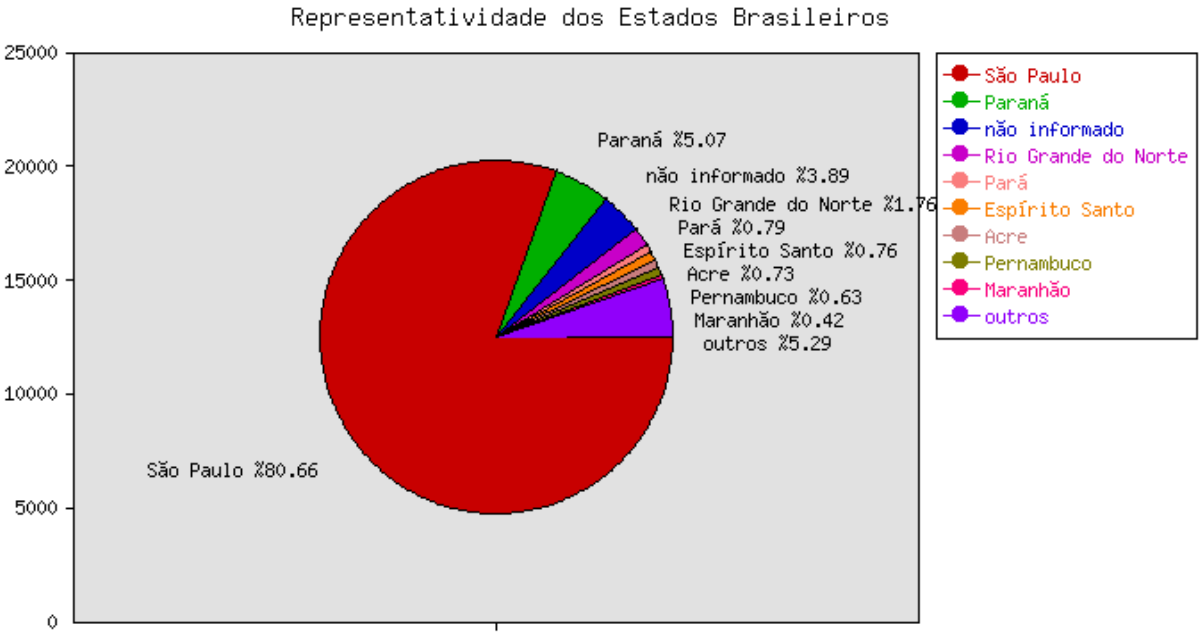
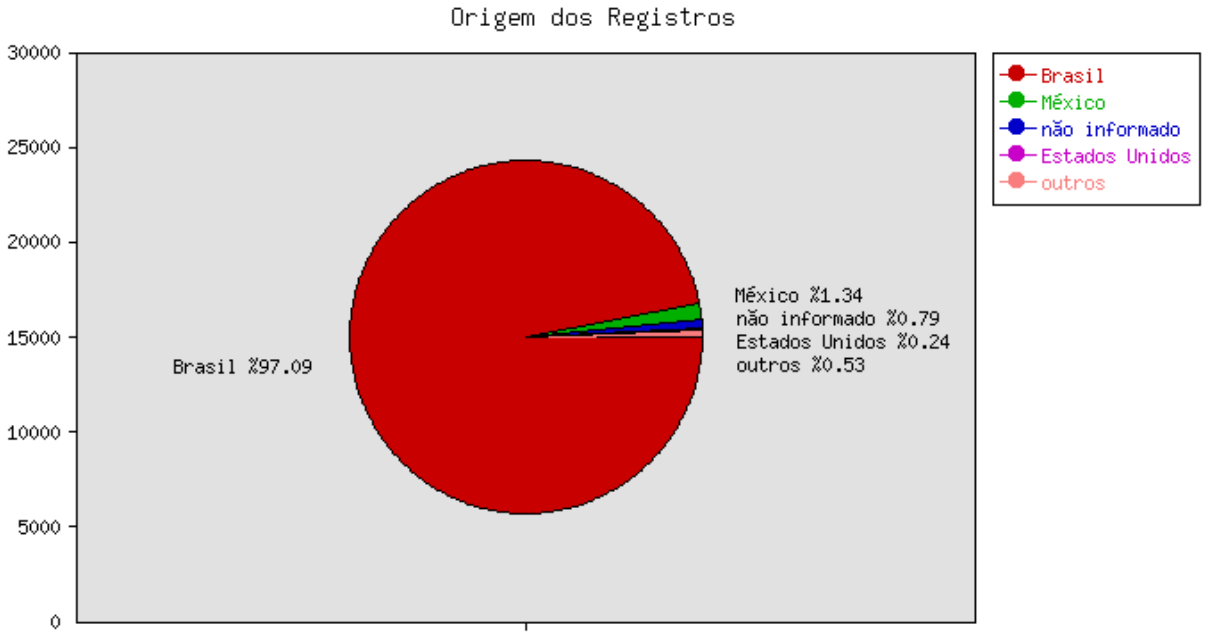


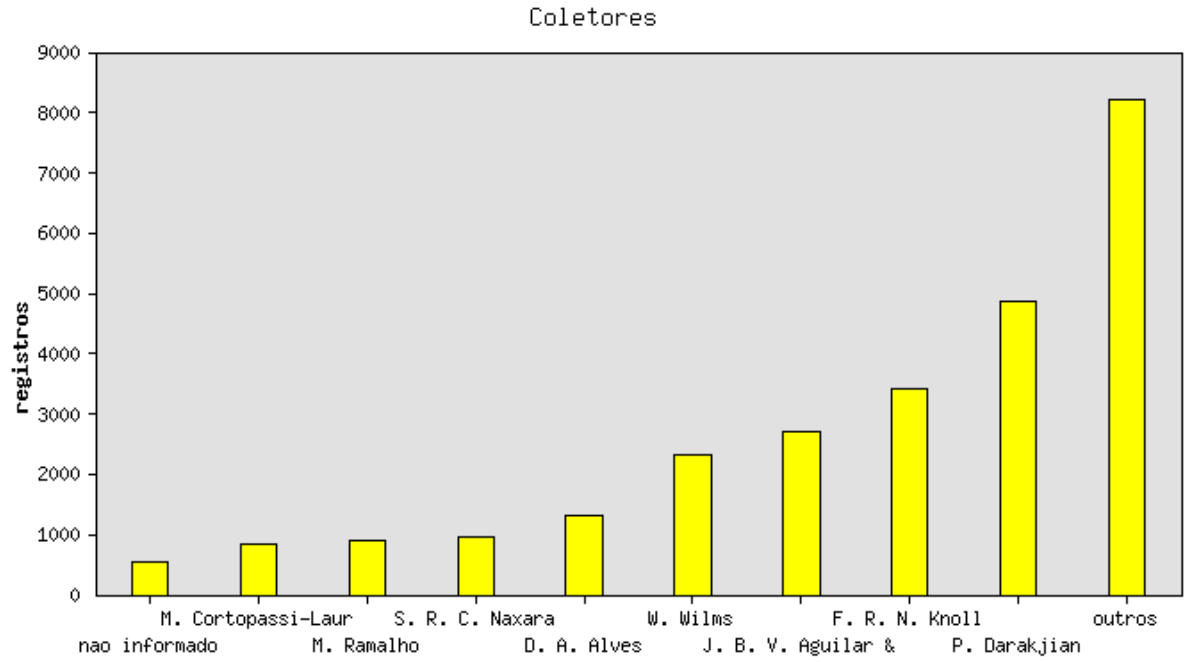
Figura 16. Perfil cadastral referente às coleções participantes da rede *species*Link

Tem-se também o perfil da coleção baseado apenas nos dados on-line. Esse perfil é apresentado em gráficos produzidos dinâmicamente. A título de exemplo apresentamos o perfil do CEPANN - Coleção Entomológica Paulo Nogueira-Neto - IB/USP. Os gráficos gerados a partir dos registros on-line apresentam as 10 famílias mais citadas, os 10 estados brasileiros mais coletados, os países de origem dos registros e os 10 coletores mais citados.


Representatividade das Famílias
registros
25000
20000
15000
10000
5000
0
Leiodidae
Phoridae
Meloidae
Andrenidae
Colletidae
Megachilidae
Vespidae
não informada
Anthophoridae
Halictidae
Apidae
outros


Representatividade dos Estados Brasileiros
25000
20000
15000
10000
5000
0
São Paulo %80.66
Paraná %5.07
não informado %3.89
Rio Grande do Norte %1.76
Pará %0.79
Espírito Santo %0.76
Acre %0.73
Pernambuco %0.63
Maranhão %0.42
outros %5.29
São Paulo
Paraná
não informado
Rio Grande do Norte
Pará
Espírito Santo
Acre
Pernambuco
Maranhão
outros





Esse conjunto de ferramentas têm motivado uma maior participação das coleções na rede *species*Link.

6.4. OpenModeller: Desenvolvimento de um Ambiente Computacional para Modelagem

Além do desenvolvimento da arquitetura da rede e do apoio às coleções para que elas digitassem os seus dados e se conectassem à rede, o projeto financiado pela Fapesp também propiciou o desenvolvimento de modelos de distribuição potencial de espécies baseado em seus nichos ecológicos. A idéia era mostrar desde o início dos trabalhos a importância de compartilhar dados. Vários modelos foram gerados em colaboração com outras equipes e dessa experiência nasceu a necessidade de estudar o desenvolvimento de um ambiente computacional para modelagem que facilitasse e agilizasse o trabalho do pesquisador. (Siqueira & Peterson 2003; Thomas et. al. 2004; Cameron et. al 2004 e Chapman et. al. 2005).

A geração de mapas de distribuição potencial de espécies é uma área inerentemente multidisciplinar, envolvendo geo-processamento, algoritmos de modelagem de distribuição de espécies, com conceitos matemáticos e estatísticos, além do conhecimento biológico e ecológico.

A geração dos mapas de distribuição também é um procedimento dentro de um SIG (Sistema de Informação Geográfico). Sua aplicação resulta em um mapa geo-referenciado que pode depois precisar ser analisado utilizando ferramentas comuns aos SIG, tais como cálculo de áreas, visualização conjunta com outros mapas, aplicação de interseção ou união com outros mapas, etc.

O CRIA está desenvolvendo o projeto *openModeller*[32] como uma biblioteca computacional de código aberto (*open source*). Usuários das diversas áreas de conhecimento poderão contribuir com o desenvolvimento do projeto e com a avaliação de seus resultados. O projeto recebeu o apoio da Fapesp e está sendo desenvolvido como uma parceria entre o CRIA, a Politécnica da USP e o INPE.

A idéia central é que na condição de biblioteca computacional, o *openModeller* pode ser facilmente integrado a outros aplicativos (ex: *plug-in* de um SIG) ou pode simplesmente servir de núcleo para uma interface de linha de comando, gráfica, web ou via web services.

Outra característica importante do *openModeller* é sua estrutura de plug-ins para os algoritmos de modelagem. O código que implementa o algoritmo deve seguir uma interface padrão (simples) que permita sua utilização pela biblioteca. Desta forma, o desenvolvedor do algoritmo pode se concentrar apenas nos problemas relativos ao próprio algoritmo sem se preocupar com os aspectos de leitura e escrita de dados, amostragem de pontos, transformações entre sistemas de coordenadas e projeções distintas, além do casamento entre mapas de escalas, dimensões e regiões distintas.

[32] http://openModeller.sf.net

A figura 17 a seguir ilustra a arquitetura básica do *openModeller* e suas relações com as interfaces e algoritmos.

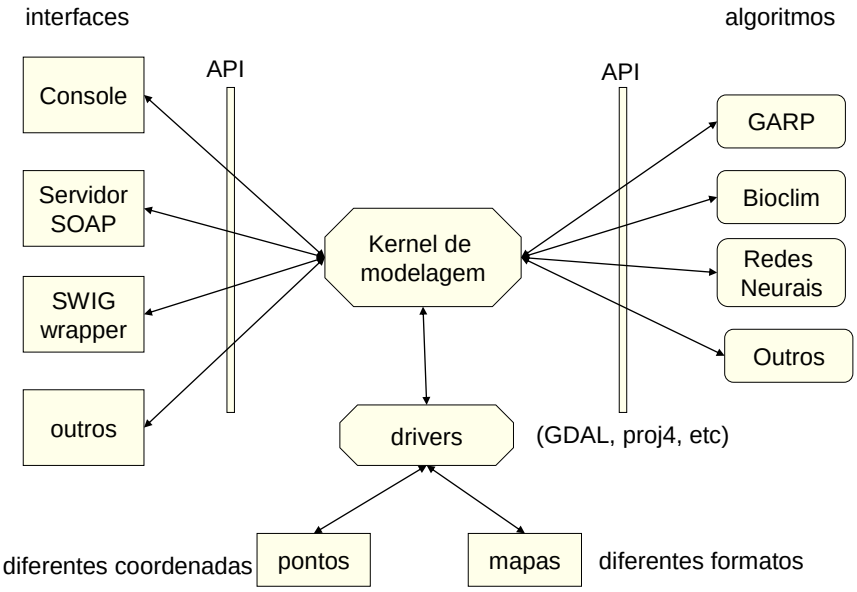


Figura 17. Arquitetura básica do openModeller

Resumidamente, o processo de modelagem segue os seguintes passos:

- Especificação dos parâmetros (dados de entrada, formato da saída, algoritmo de modelagem)
- Leitura dos dados ambientais (cada mapa geo-referenciado representa os valores de uma variável ambiental distinta)
- Leitura dos dados biológicos (pontos geo-referenciados de ocorrência da espécie em questão)
- Cruzamento dos dados biológicos com os ambientais
- Geração do modelo utilizando o algoritmo especificado.
- Projeção do modelo gerando o mapa de distribuição de espécies (MDE).

O openModeller está sendo desenvolvido no Source Forge[33] e o seu código fonte está disponível sob a licença GNU *General Public License*[34]. Também estão disponíveis versões binárias que podem ser utilizadas por desenvolvedores de algoritmos. O intuito de utilizar um repositório como o Source Forge foi o de facilitar o desenvolvimento colaborativo. O *openModeller* conta com a colaboração de 6 desenvolvedores, sendo um do CRIA e cinco de fora do Brasil[35].

[33] http://sourceforge.net/projects/openModeller

[34] http://www.gnu.org/copyleft/gpl.html

[35] http://openModeller.sourceforge.net/index.php?option=com_contact&Itemid=3

O código está sendo desenvolvido utilizando a linguagem C++ ANSI, o que torna o código facilmente portável. O desenvolvimento está sendo feito em Linux, porém pode ser compilado para sistemas operacionais UNIX em geral e para Windows. A portabilidade para MacOs está em fase de desenvolvimento. Apesar do código estar em C++, interfaces para as linguagens Python e Java também podem ser geradas através da ferramenta Swig[36]. Hoje existem também protótipos para uma interface SOAP e uma interface Web, além de uma interface gráfica multiplataforma que funciona como plugin do SIG Quantum GIS[37], também de código aberto.

A biblioteca *openModeller* está sendo também utilizada pelo projeto "Biodiversity World" (BDWorld[38]), pelo projeto "Science Environment for Ecological Knowledge" (SEEK[39]), e pelo projeto “Biological Terrorism Risk Assessment” (BTRA[40])

A versão atual do *openModeller* conta com cinco algoritmos e três variações:

- Bioclim (implementado por Mauro Muñoz - CRIA)
- Bioclim distance (implementado por Mauro Muñoz - CRIA)
- Climate Space Model - Broken-Stick (implementado por Tim Sutton - BDWorld)
- Climate Space Model - Kaiser-Gutman (implementado por Tim Sutton - BDWorld)
- Distance to average (implementado por Mauro Muñoz - CRIA)
- GARP: Genetic Algorithm for Rule Set Production (implementado por Ricardo Scachetti Pereira – Universidade de Kansas)
- GARP with Best Subsets Procedure (implementado por Ricardo Scachetti Pereira – Universidade de Kansas)
- Minimum distance (implementado por Mauro Muñoz - CRIA)

Os detalhes de cada algoritmo estão disponíveis na página do projeto[41].

7. Infra-estrutura

É importante avaliar a infra-estrutura necessária para o desenvolvimento e manutenção de sistemas de informação de acesso livre e aberto do porte daqueles desenvolvidos para o Programa Biota/Fapesp e para o Programa de Biotecnologia do MCT.

7.1. Hardware

O CPD do CRIA é composto por dois servidores modelo PowerEdge 6600, cada um com 4 processadores Intel Pentium III Xeon, 2GB de memória RAM por processador e capacidade de disco total de 1.5 TB e uma unidade de *backup* em fita modelo Dell/EMC com capacidade para 20 fitas DLT com 100GB de capacidade cada uma. Além dos dois servidores principais, o CRIA possui mais quatro servidores menores, utilizados para geração de mapas, recuperação de informação, gerenciamento de serviços de impressão e testes da equipe de desenvolvimento.

---

[36] http://www.swig.org

[37] http://qgis.org

[38] http://www.bdworld.org

[39] http://seek.ecoinformatics.org

[40] http://www.specifysoftware.org/Informatics/informaticsbtra/

[41] http://openModeller.sourceforge.net/index.php?option=content&task=category&sectionid=3&id=9&Itemid=39

Todo o CRIA é interligado por cabeamento de dados estruturado, suportando uma capacidade de transferência de até 100 Mbps. Os servidores principais são interligados por uma sub-rede de fibra óptica que suporta transferências de até 1 Gbps.

O CRIA possui uma conexão de dados com o nó de Campinas da rede ANSP ("*Academic Network of São Paulo*") formada por um *link* de fibra óptica entre o prédio do CRIA e o Centro de Computação da UNICAMP com velocidade de 1 Gbps. O nó da UNICAMP está conectado à FAPESP por uma conexão de 155 Mbps. Essa conexão entre o CRIA e a rede ANSP é controlada por dois roteadores Foundry de última geração. A rede interna do CRIA é protegida por um *firewall* instalado nos roteadores e por *firewalls* locais a cada um dos servidores. Todo o tráfego de rede para dentro e para fora do CRIA é examinado para prevenir a entrada de *virii* de computador.

O CPD do CRIA também possui um aparelho de ar condicionado AirSplit 24000BTU e um no-break modelo Prestige 6000.

7.2. Software

Todos os sistemas e ferramentas desenvolvidos no CRIA funcionam sobre o sistema operacional Linux usando apenas ferramentas de software livre. Os dados são armazenados em um sistema gerenciador de bancos de dados PostgreSQL e o software foi desenvolvido em linguagens PHP, Perl e Java. Como protocolos de transferência de dados são utilizados HTTP, SOAP e XML. Todo o sistema faz uso do software Apache como servidor de páginas web.

A equipe de suporte monitora diariamente sites de desenvolvedores de *software* e listas de discussão de falhas de segurança para avaliar quando e se um software deve ou não ser atualizado. O *software* dos servidores do CRIA (inclusive o próprio sistema operacional) é atualizado somente quando uma nova versão possui uma característica útil ao projeto sendo desenvolvido, quando a equipe de suporte considera que a instalação de uma nova versão tornará o sistema mais eficiente e principalmente quando é descoberta alguma falha de segurança em algum programa ou parte do sistema operacional.

O sistema de verificação de vírus e invasões de *hackers* é atualizado diariamente. O *software* dos *desktops* dos pesquisadores é atualizado semanalmente, de acordo com as recomendações dos fabricantes, com o objetivo de eliminar falhas de execução e brechas de segurança.

É feito um backup diário dos dados em disco. Semanalmente é feito um backup completo do sistema em fita. Todo mês uma cópia de segurança da fita com o back-up completo (sistema e dados) é armazenada nas dependências da Embrapa Informática Agropecuária (CNPTIA).

8. Sustentabilidade

Uma das conseqüências da dependência crescente por informações nas diferentes áreas de conhecimento de forma integrada e interoperável é a necessidade de maiores investimentos no gerenciamento e preservação de dados. Não deve ser da responsabilidade do cientista desenvolver e manter sistemas de acesso público a dados. Seus esforços e competência devem estar concentrados em trabalhos de análise, interpretação e síntese. A tarefa de desenvolver e manter sistemas de informação complexos é trabalho de profissionais cuja função também é desenvolver estratégias para o manejo de dados e informações para as próximas décadas ou séculos.

Cabe, portanto a um centro de informação:

- Desenvolver um projeto de planejamento de longo prazo;

- Atuar de forma integrada com a comunidade científica e se basear na demanda e nas orientações desta comunidade para as decisões sobre arquivamento de longo prazo;
- Obter suporte financeiro de longo prazo para o centro de dados e operações de manutenção e arquivo;
- Contar com equipe qualificada capaz de proceder atualizações permanentes de dados, *software* e *hardware*; além de
- Trabalhar de forma colaborativa com a comunidade científica local, nacional e internacional.

O CRIA atua exatamente nesse nicho e está conseguindo realizar um trabalho relevante para a sociedade, contribuindo para aumentar o acesso livre, aberto e gratuito a dados e informações. O grande desafio é sua sustentabilidade financeira.

O CRIA é uma Organização da Sociedade Civil de Interesse Público (OSCIP). Diferentemente de uma instituição pública, o CRIA não tem um aporte fixo de recursos para a sua manutenção. Diferentemente também de uma empresa privada, não dispõe de um produto comercial "vendável" que possa garantir a sua sobrevivência. O CRIA atualmente recebe apoio "por projetos" ou "por serviços". O apoio "por projeto", além de ser de curto ou no máximo de médio prazo, esbarra em um grande entrave da maioria das agências financiadoras, que é a inexistência de recursos para pagamento de pessoal.

Ainda, pelo tipo de atividade que exerce, praticamente todo projeto e prestação de serviços deixa mais um sistema de informação ou um banco de dados que precisa ser mantido, mesmo quando os recursos já acabaram. Essa manutenção, apesar do seu trabalho ser reconhecido nacional e internacionalmente, depende da aprovação constante de novos projetos que, por sua vez criam novas demandas da equipe e da infra-estrutura.

Projetos são importantes e até essenciais para estudos específicos mas a verdadeira inovação vem com um financiamento estável e de longo prazo. Um estudo realizado pelo National Science Board da National Science Foundation (NSB, 2005) recomenda que o apoio ao dado, deve ser permanente e o apoio ao gestor do sistema de informação deve ser de longo prazo com avaliações periódicas.

Torna-se premente, portanto, um estudo de um outro modelo de financiamento das atividades de instituições gestoras de sistemas de informação, sejam elas públicas ou de interesse público.

9. Referências

Camaeron, A., Thomas, C.D., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O. & Williams, S.E.. 2004. Will climate change catch us off guard? 2004. **Conservation In Practice** V5(2):28-30.

Canhos, D.A.L. (coordenação), Canhos, V.P., Souza, S., Siqueira, M.F., Muñoz, M., Giovanni, R., Marino, A., Koch, I., Fonseca, R.L., Umino, C.Y., Cruz, B. e Albano, A.P.S. Sistema de Informação Distribuído para Coleções Biológicas:a Integração do *Species Analyst* e *Sin*Biota. Relatório Técnico Anual. Fapesp. Outubro de 2004. splink.cria.org.br/docs/outubro2004.pdf

Chapman, A. D., M. E. S. Muñoz, and I. Koch. 2005. Environmental information: Placing environmental phenomena in an ecological and environmental context. Biodiversity Informatics 2:24-41.

Chapman, A.D. (1999). Quality control and validation of point-sourced environmental resource data pp.409-418 in Lowell, K. and Jaton, A. (eds). Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources. Chelsea, Michigan: Ann Arbor Press. 455pp

Chapman, A.D. (2004). Environmental Data Cleaning Tools – A Discussion Paper. http://splink.cria.org.br/docs/appendix_i.pdf.

Chapman, A.D. (2004). Environmental Data Quality – A Discussion Paper. http://splink.cria.org.br/docs/appendix_h.pdf.

CRIA. 2005. Esquema de conexão da rede speciesLink. http://splink.cria.org.br/manager/esquema36.pdf. [12 de maio de 2005]

Döring, M. e Giovanni, R. 2004. GBIF Data Access and Database Interoperability – A unified protocol for search and retrieval of distributed data. http://www.cria.org.br/protocols/newprotocol.pdf.

National Science Board. Draft Report: Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. NSB-05-40. March 30, 2005. http://www.nsf.gov/nsb/meetings/2005/LLDDC_draftreport.pdf

Siqueira, M. F., Peterson, A. T. 2003. Consequences of global climate change for geographic distributions of cerrado tree species. **Biota Neotropica**. , v.3, n.2, http://www.biotaneotropica.org.br/v3n2/pt/download?article+BN00803022003+item

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C. Erasmus, B. F. N., Siqueira M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T. Phillips, O. L. & Williams, S. E. 2004. Extintion risk from climate change. **Nature** 427(8)145-148.