Strain Data Networks

The *species*Link and SIColNet experiences

Dora Ann Lange Canhos
Centro de Referência em Informação Ambiental - CRIA

Centro de Referência em Informação Ambiental

CRIA (Reference Center on Environmental Information) a not-for-profit, non-government organization.

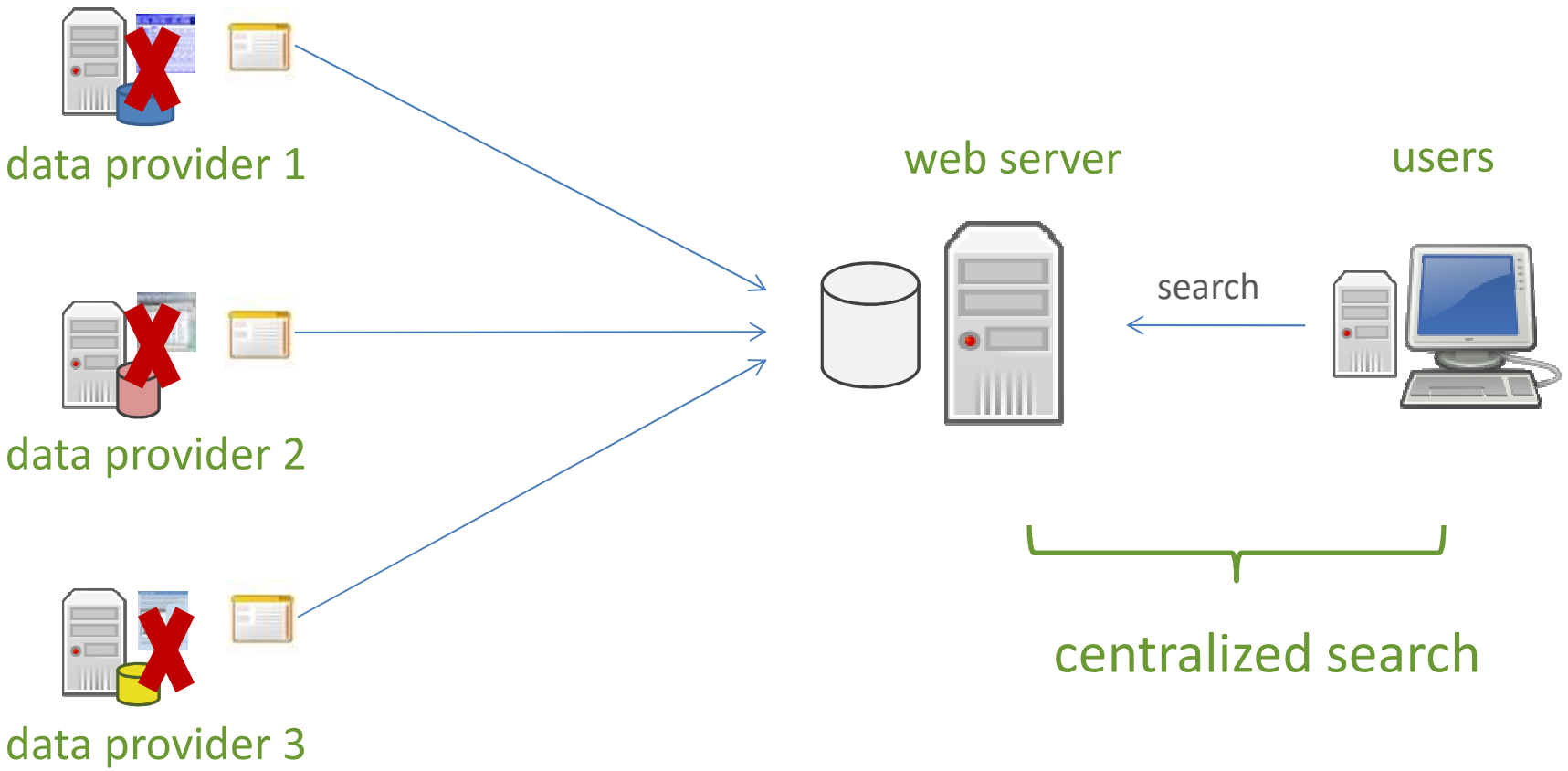
Its aim is to contribute towards a more sustainable use of Brazil's biodiversity through the dissemination of high quality data and information generated by the scientific community

Focus for the GBRCN information system

- Linking existing systems – information exchange
- Quality – data validation
- Data content
- Data usability

Strategies for data integration

1) Same software and database used by all providers



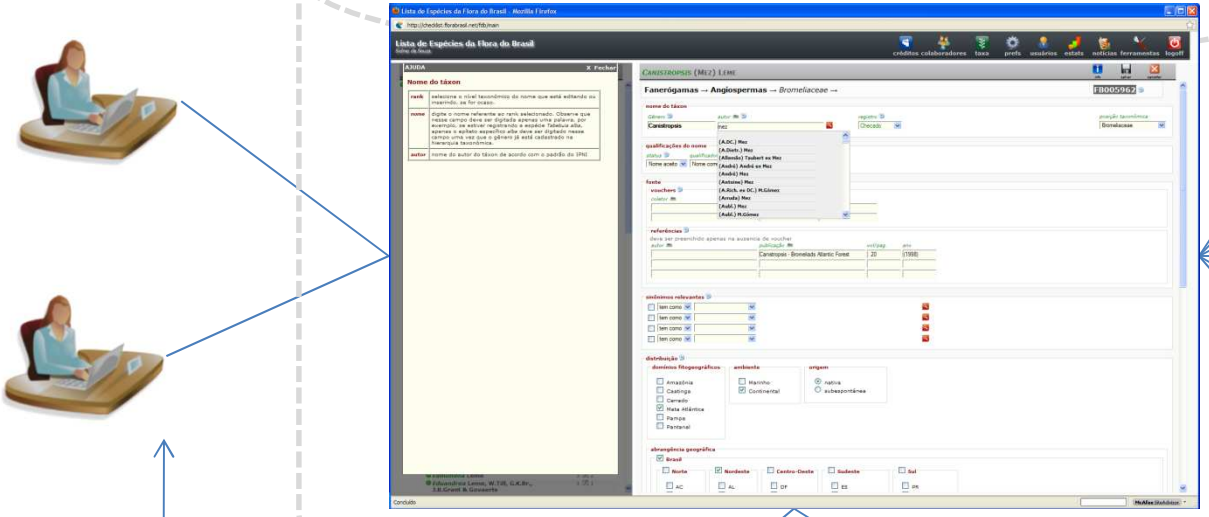
Slide: Renato de Giovanni (CRIA)

Strategies for data integration

1) Same software and database used by all providers

- ✓ Interesting solution if all providers agree to use the same system:
  - ✓ Improvements benefit all participants.
  - ✓ Shared costs.
- ✓ Good performance (although queries are run in the production database).
- Lack of freedom to make custom adjustments.
- Very difficult to accomplish if providers are already using their own management software (sometimes developed with considerable effort).

Development team



Specialists (>400)

HELP

HELP

New Developments

Coordinating group

lista de espécies
flora do brasil

development CRIA

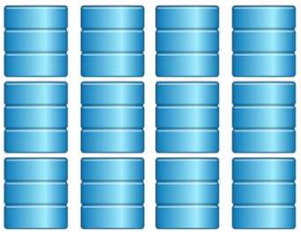
- Importing data from existing lists
- Maintenance, correcting bugs
- New implementations
- Support to the coordination

coordination JBRJ

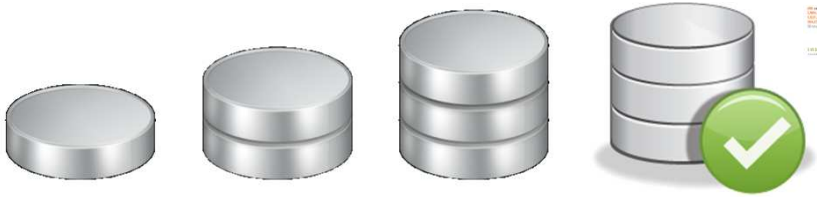
- User control
- Global corrections
- logs and controls
- Support to taxonomists

taxonomists

- data cleaning interface
- Statistics interface
- Editing interface
- External resources

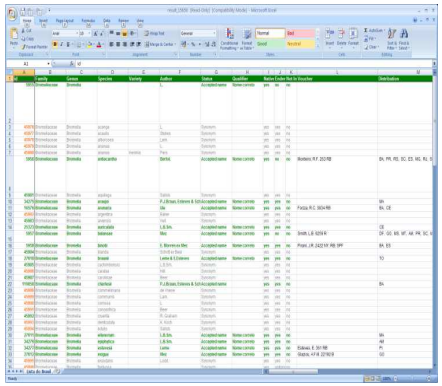


backups, backups, backups, backups ...

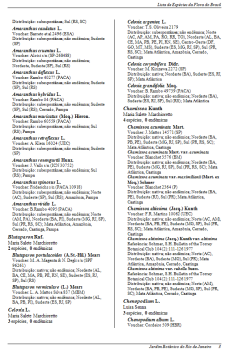


Web interface

saída xls planilha



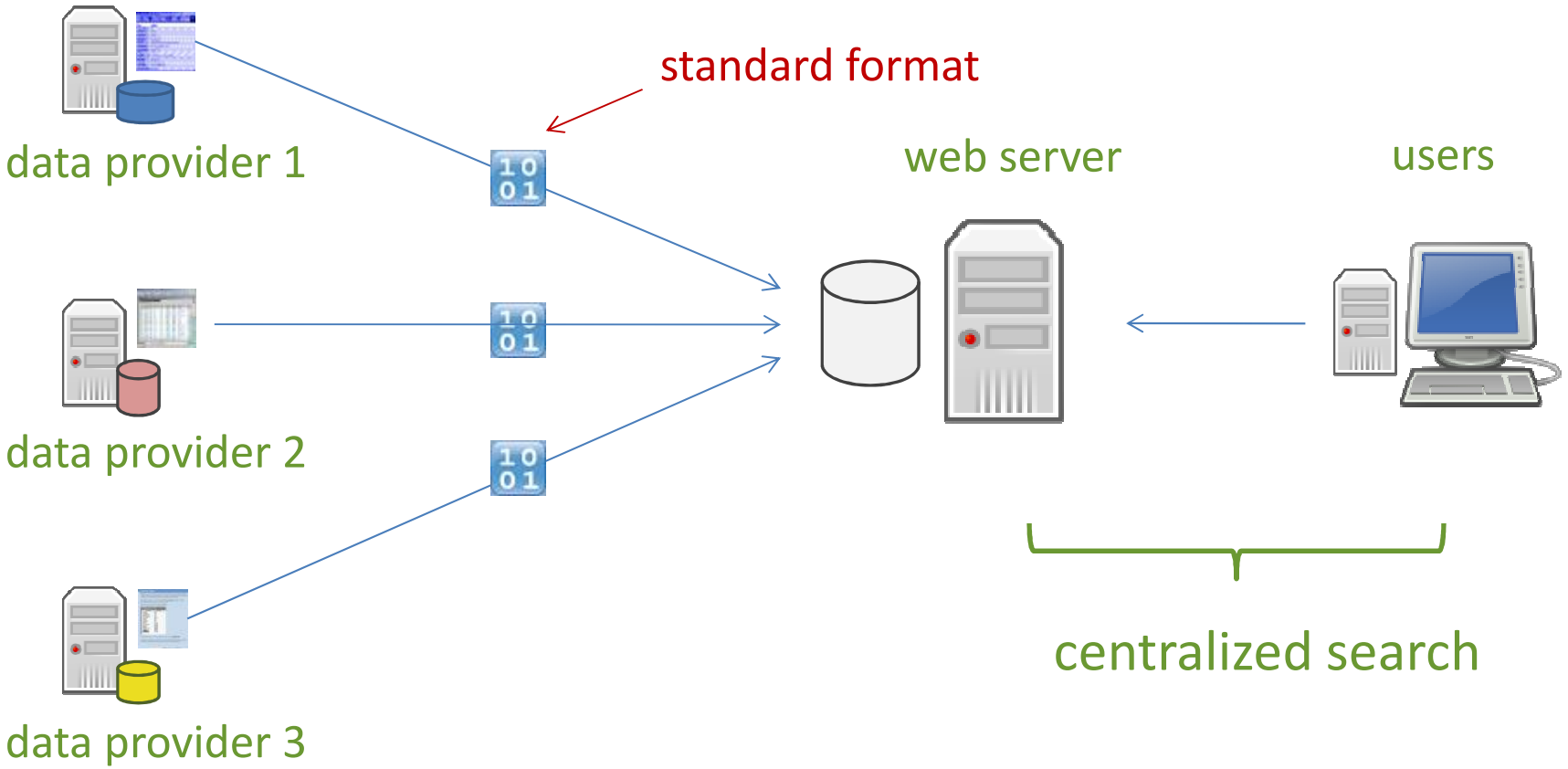
saída rtf para impressão



mapas de distribuição

Strategies for data integration

2) Periodically export data to a central database



Strategies for data integration

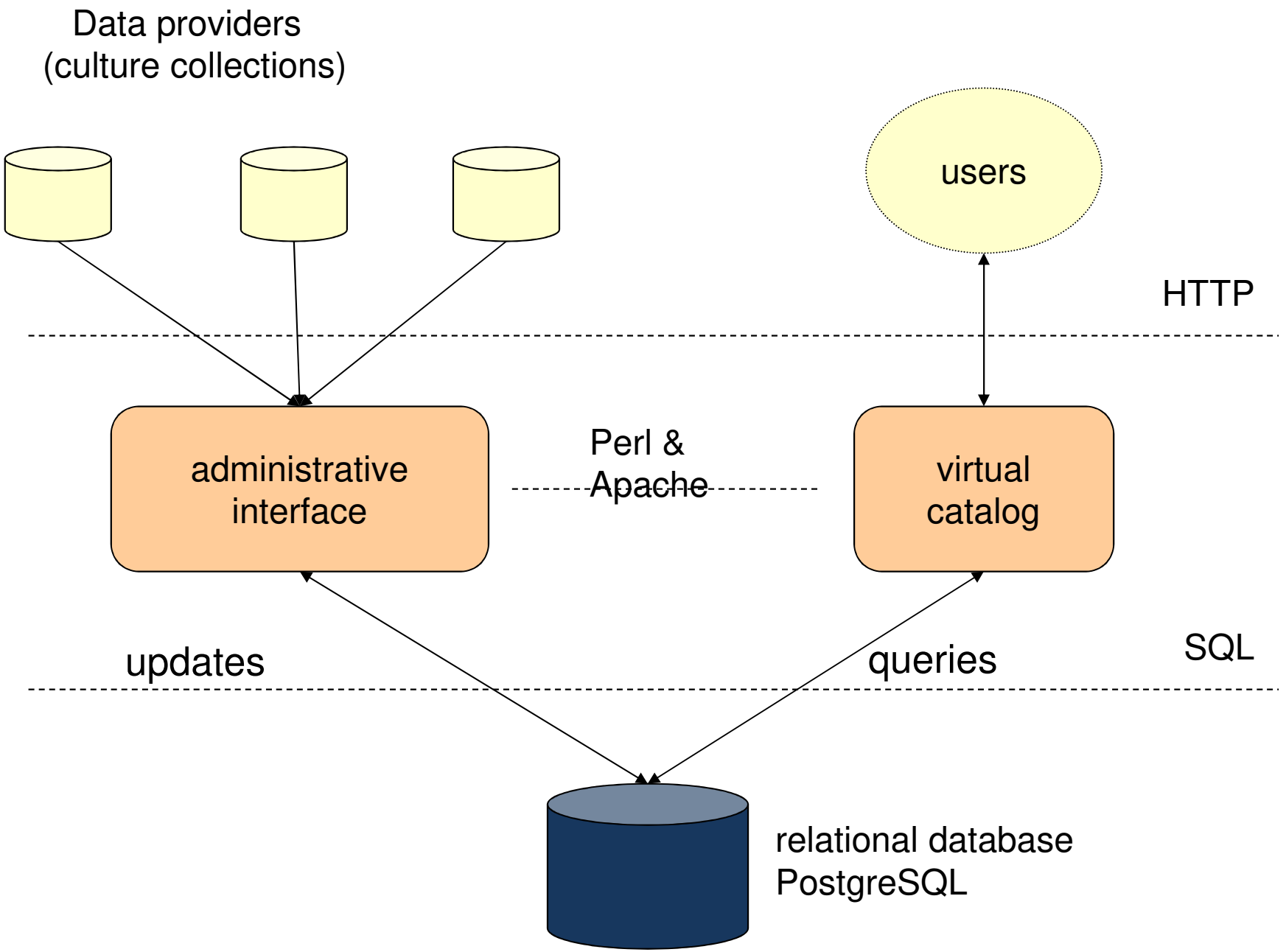
2) Periodically export data to a central database

Examples:

- **C**ommon **A**ccess to **B**iological **R**esources and **I**nformation.
- Began in 1999.
- 28 catalogues from European institutions (>100K records).

SICol

- 1st phase of the Brazilian network


Data providers
(culture collections)
users
HTTP
administrative
interface
Perl &
Apache
virtual
catalog
updates
queries
SQL
relational database
PostgreSQL

Strategies for data integration

2) Data providers periodically export data to a central database

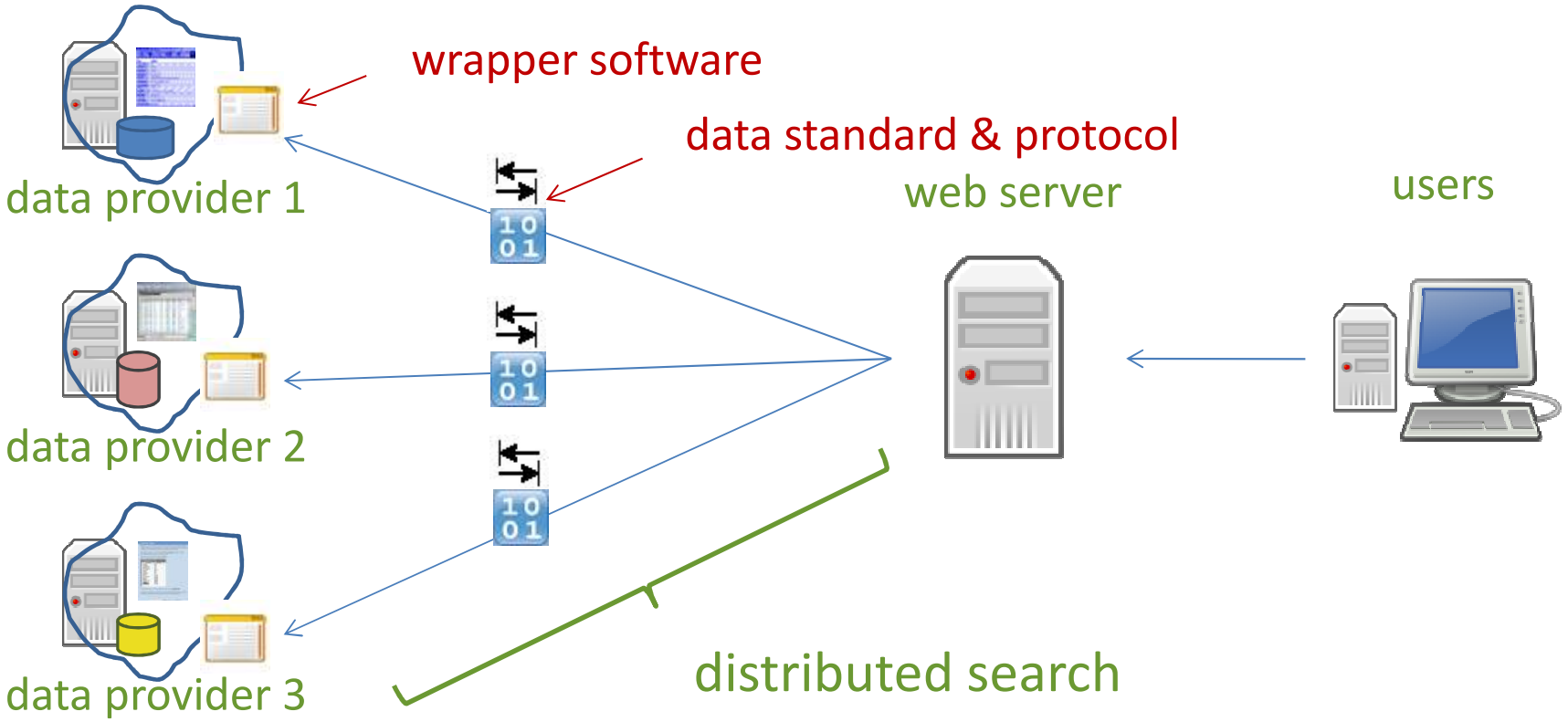
✓ Good performance.

✓ Easier to implement.

- Queries are performed on potentially non current data.
- Onus on providers to transform data into a common format and periodically export it.
- Experience with SICol: no updates

Strategies for data integration

3) Real time distributed queries



Strategies for data integration

3) Real time distributed queries

Examples:

- 1998 - 2003.
- North America.
- MaNIS, HerpNET, ORNIS & FishNet.

REMIB

Red Mundial de Información sobre Biodiversidad

- Started in 1998.
- Mexico.

AVH

Australia's Virtual Herbarium

- Started in 2000.
- 9 major herbaria.
- 6 million records (80% databased).

Strategies for data integration

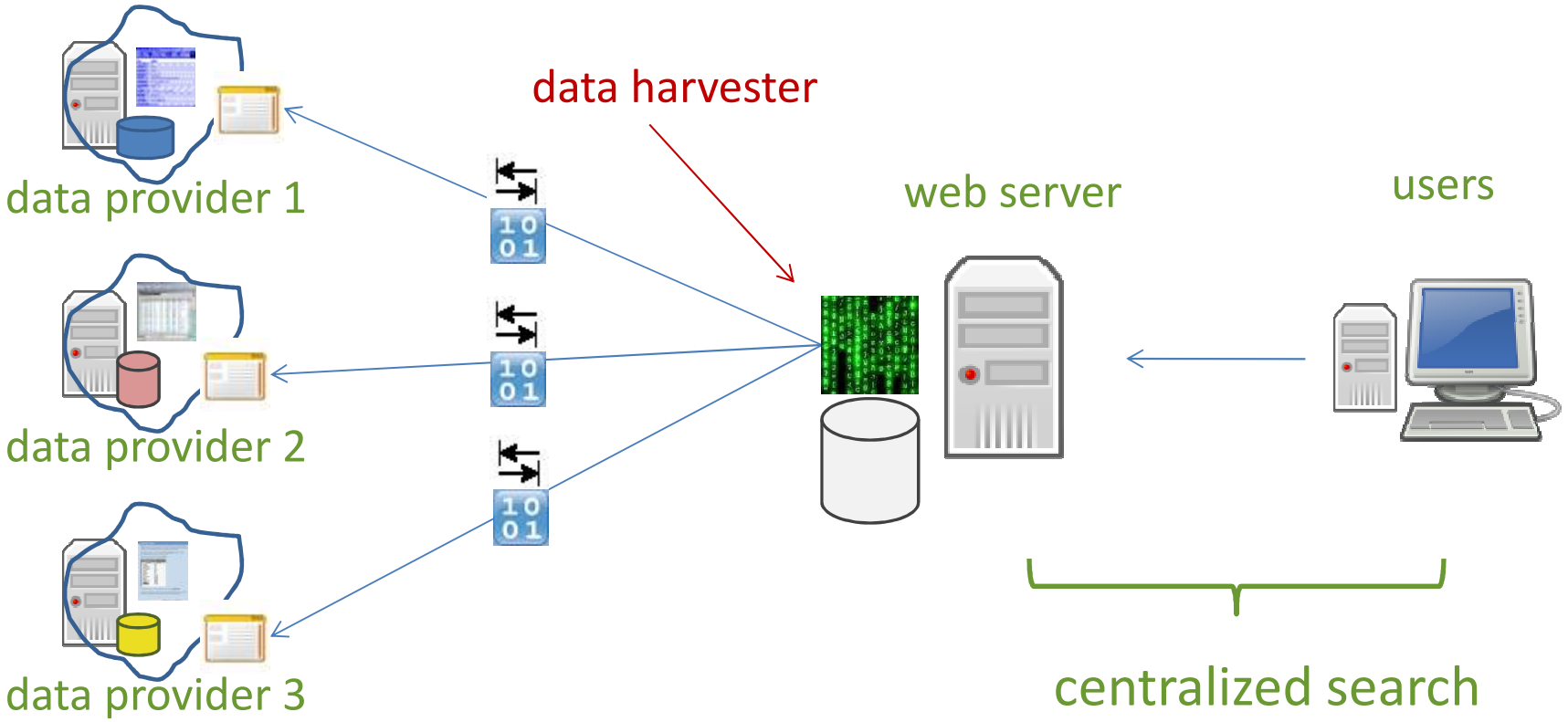
3) Real time distributed queries

- ✓ Access to current data.
- ✓ Providers have more confidence and sense of control.

- Performance and scalability bottlenecks.
  - Performance limited by the slowest data provider.
  - Servers sometimes down, network problems.
  - When data providers go offline their data become unavailable.

Strategies for data integration

4) Data harvesting



Strategies for data integration

4) Data harvesting

Examples:

OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

speciesLink

Strategies for data integration

4) Data harvesting

✓ Good performance.

- It may be necessary to define a common (minimum) field set for storing data in the central database.

- Queries are performed on potentially non current data.
- Difficult to implement if there are many protocols and data standards involved.

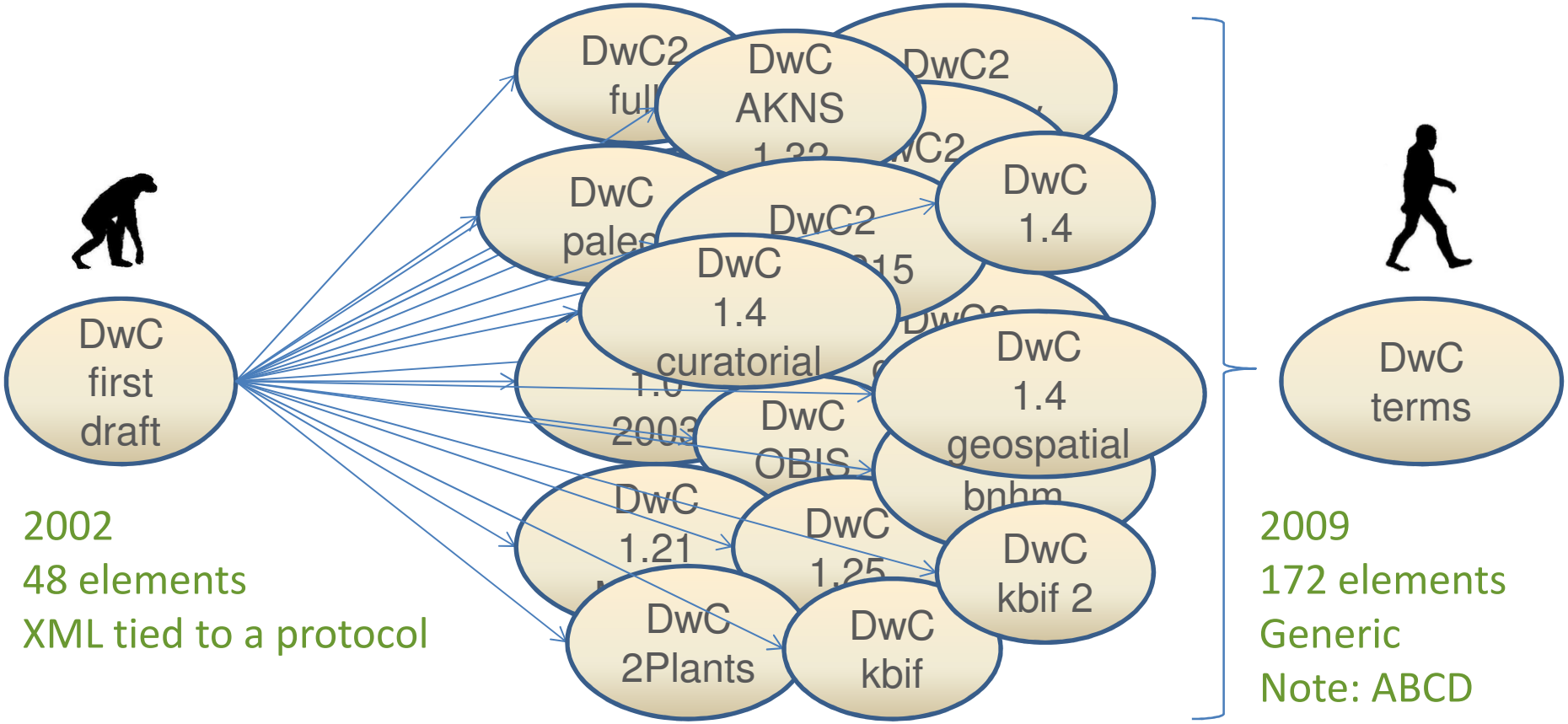
Data Model: DarwinCore

- Based on specifications developed by the DublinCore Metadata Initiative. Can be seen as an extension of it for biodiversity data.

- Its latest version consists of a glossary of terms including definitions, examples, and commentaries, including how terms:
  - are managed
  - can be used
  - can be extended for new purposes

- Designed to minimize the barriers to adoption and to maximize reusability in a variety of contexts.

On the evolution of Darwin Core



2002
48 elements
XML tied to a protocol

~20 different versions!

Standardization required!

2009
172 elements
Generic
Note: ABCD has 970 terms

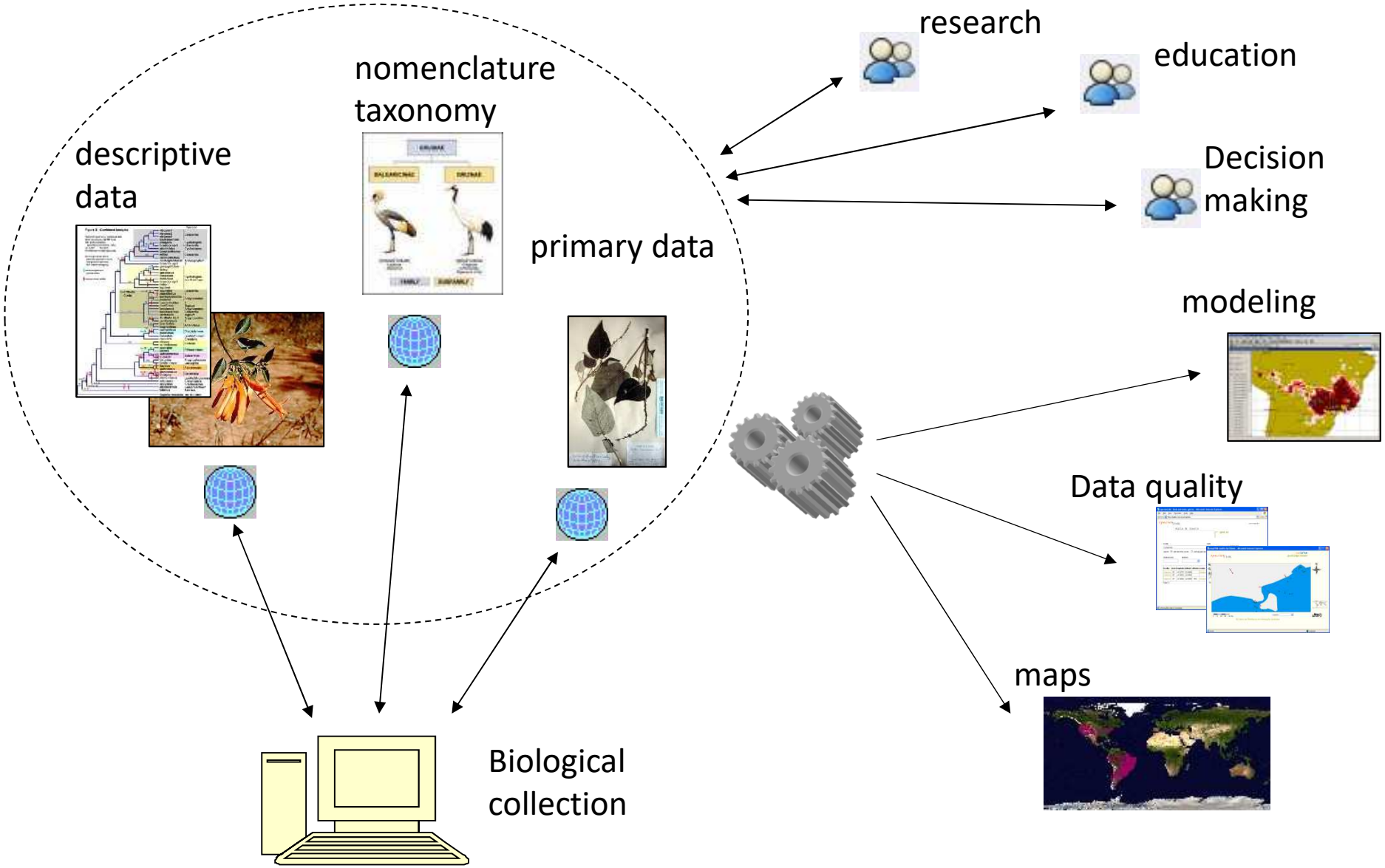
Data exchange protocol - TAPIR

- TDWG Access Protocol for Information Retrieval.
- Integrates functionality from DiGIR and BioCASe.
- Completely independent of the data being exchanged: Works with DarwinCore and ABCD.
- Official TDWG standard.
- Tools and documentation available.

Choosing standards and protocols

- Choose from existing standards whenever possible:
  - This can save you considerable time.
  - Will likely avoid interoperability issues in the future.
- Seek compatibility with other initiatives.
  - You can benefit from existing tools.
  - You may get extra functionality/data.
- Data providers are the pillars of every network:
  - Help them improve their data.
  - Ensure that data remain curated at the source.
  - Show them that data sharing promotes citation and usage, giving them credits and visibility.

The speciesLink network



Points to consider

- Biological collections in Brazil
  - A small number of “large” collections
  - A great number of important small research collections
- Average characteristics
  - Human resources: expertise in informatics (normally insufficient)
  - Equipments and installations (normally insufficient)
  - Connectivity (normally slow or unstable)

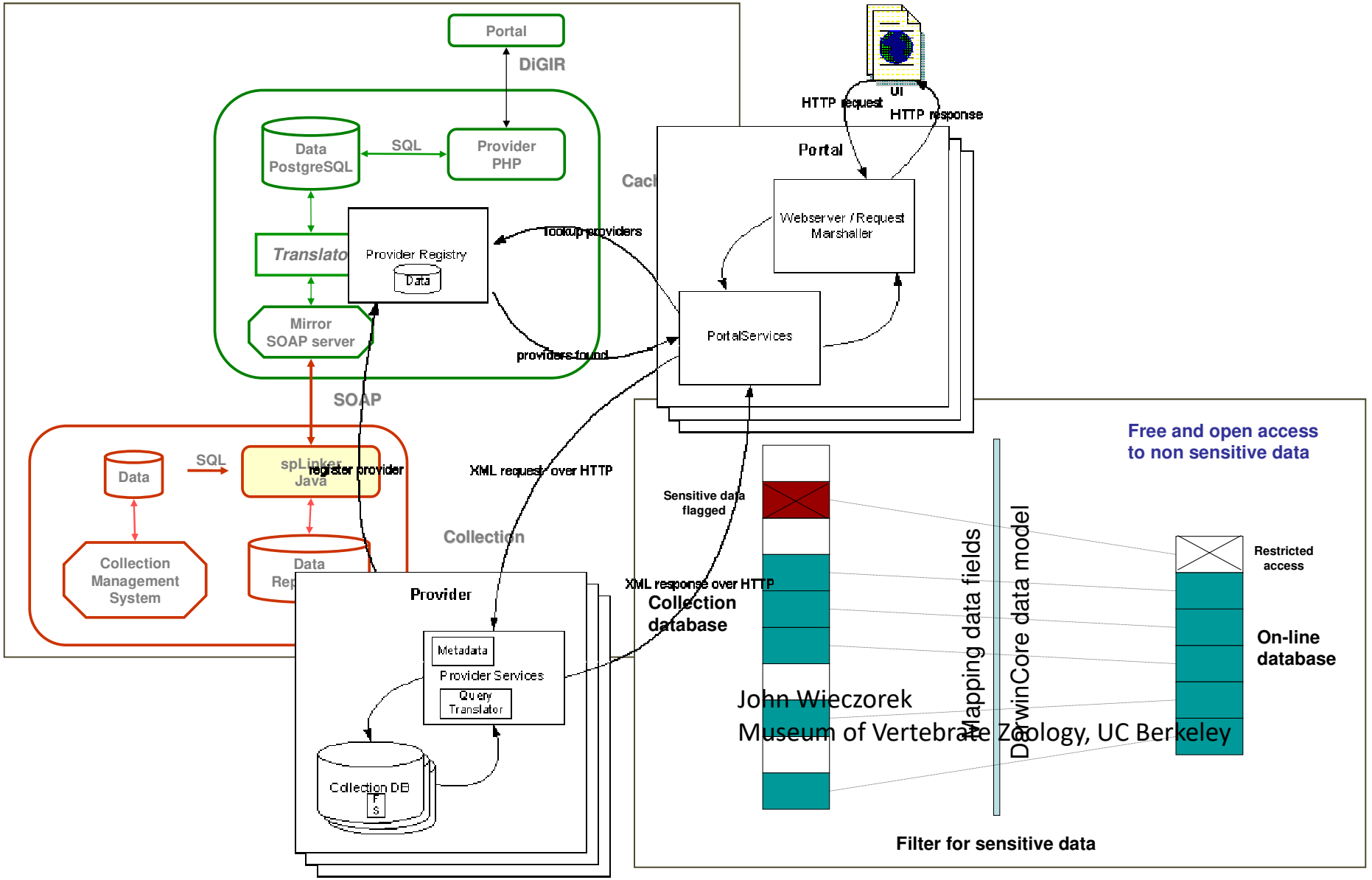
Challenges

- Integration of primary data from all taxa, from distributed collections, using different software in diverse environments
- Integrating data from collections with low and/or unstable internet connectivity, using basic hardware and no computer expertise
- Maintain full control over the data served to the network at the provider's end

Development parameters - architecture

- Collection's routine must not be altered
  - Practically any software is accepted (Excel, Access, Specify, Biota, Brahms, PostgreSQL, MySQL, …)
- Data provider must have full control over the data
  - What is sensitive data, what is open and free
  - Digitization strategy, data cleaning strategy
- Data provider must be fully acknowledged
- Connectivity problem must be overcome
- Network must be interoperable with international initiatives

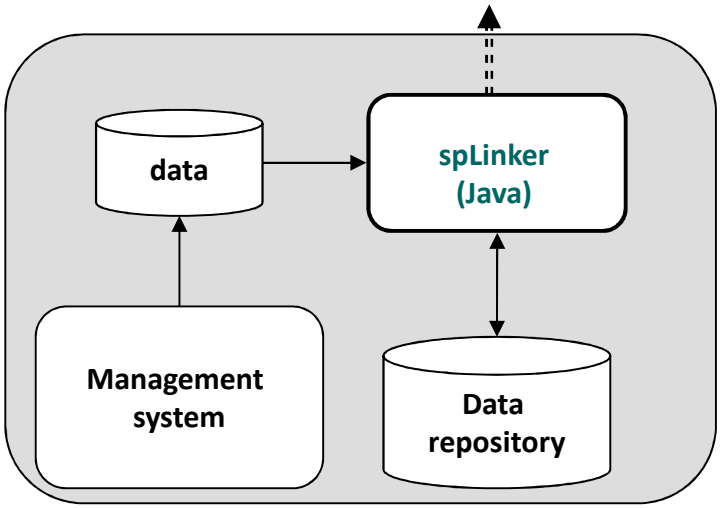
Network architecture

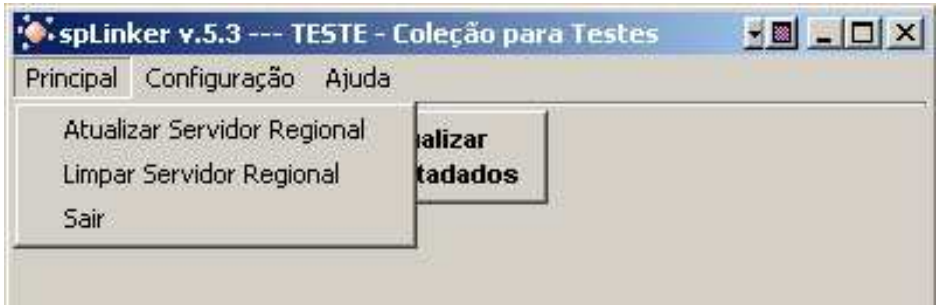


John Wieczorek
Museum of Vertebrate Zoology, UC Berkeley

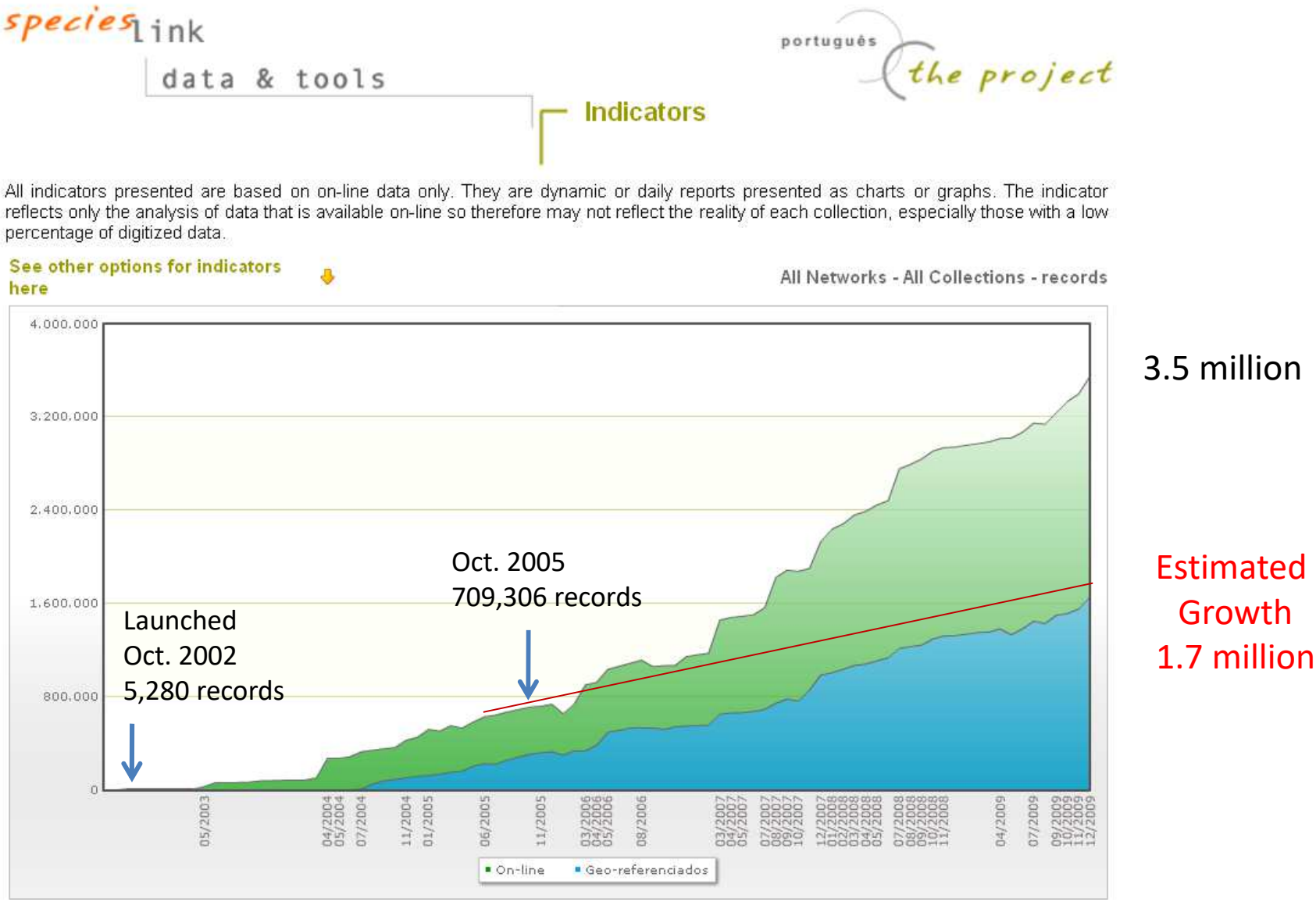
spLinker: software to send data to cache node

- Platform Independent (java)
- Connects to practically any database
- Offers full control over data
- Checks repository and only sends updates (low traffic)
- It is possible to filter sensitive data using regular expressions





The development of the speciesLink network

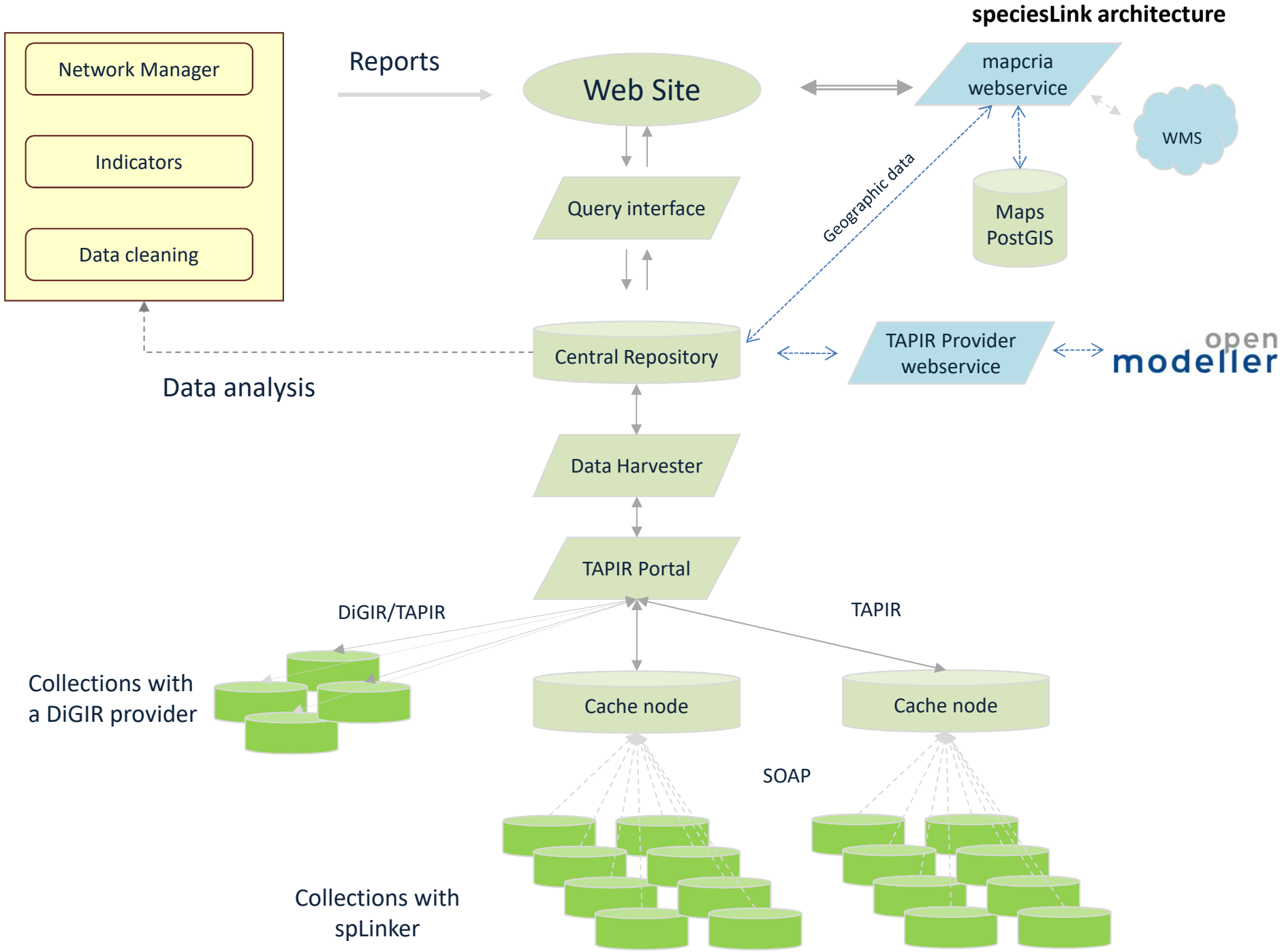


3.5 million

Estimated Growth 1.7 million

Data Sharing

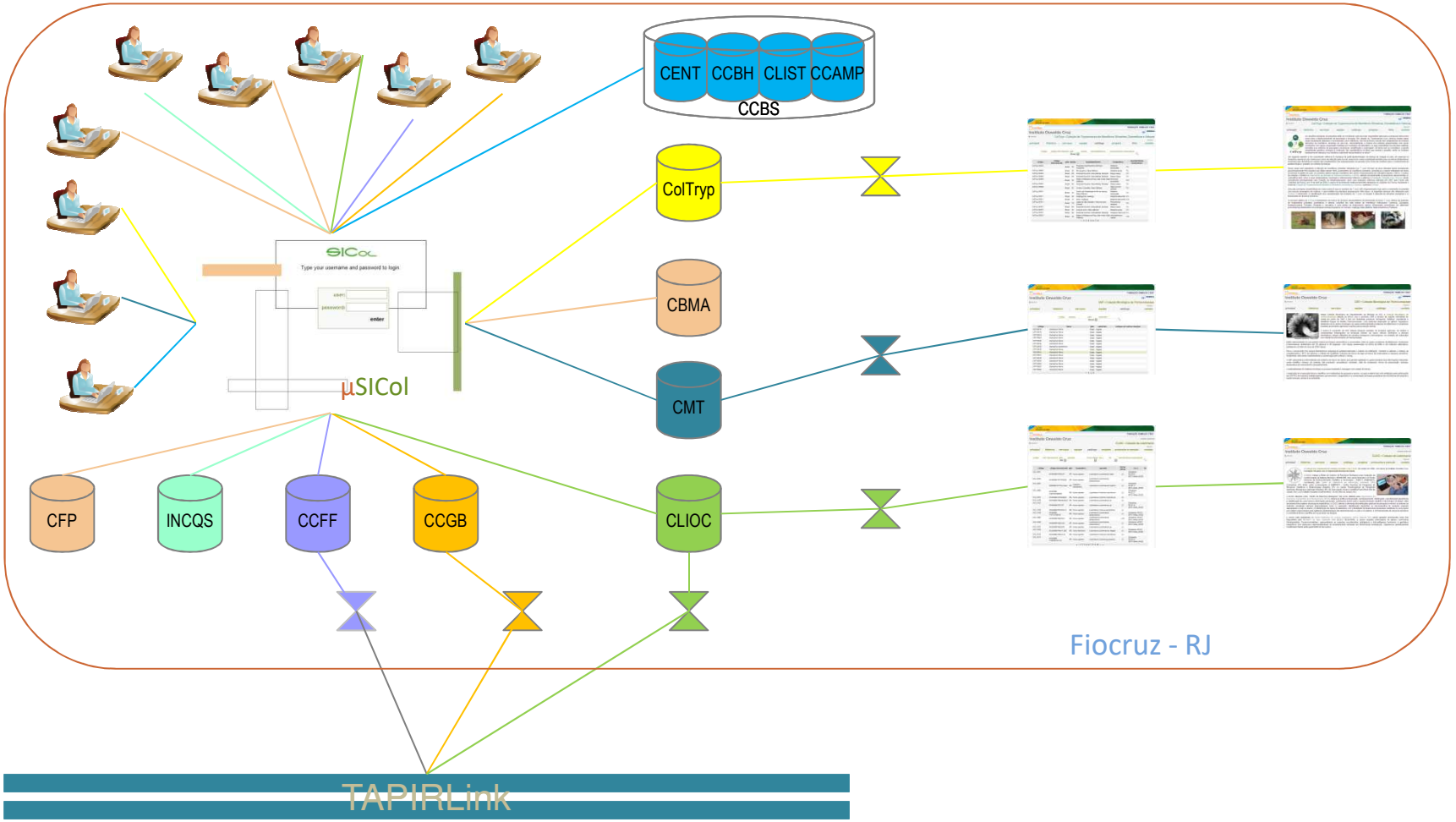
- Does not depend only on the will to share data
  - It must be planned: adequate resources, expertise, infrastructure
  - Must be organized: data models, controlled vocabulary, communication protocols, ...
  - Must be easy or at least “doable”
  - Must have a compatible data policy: free and open access to non sensitive data


speciesLink architecture
Network Manager
Indicators
Data cleaning
Reports
Web Site
mapcria webservice
WMS
Query interface
Geographic data
Maps PostGIS
Central Repository
TAPIR Provider webservice
open modeller
Data analysis
Data Harvester
TAPIR Portal
DiGIR/TAPIR
TAPIR
Collections with a DiGIR provider
Cache node
Cache node
SOAP
Collections with spLinker

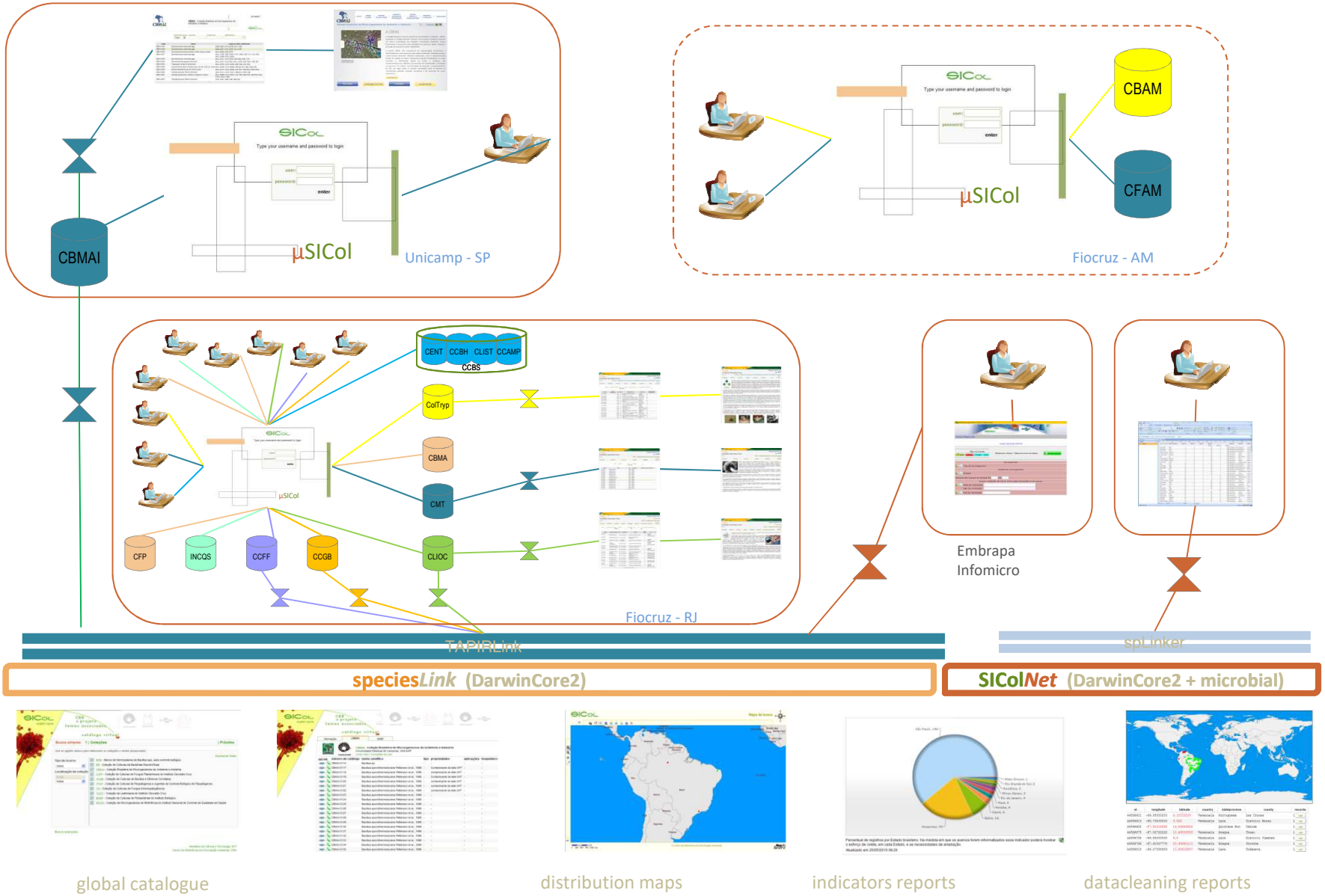
Lessons learned

- ❑ Adoption of internationally agreed standards and protocols is key
- ❑ Support unlocking and sharing of data (make it simple and easy !)
- ❑ Enable data providers to have full control of their data determining what can be openly shared and what is sensitive
- ❑ Full credit and acknowledgement to the data providers at all levels !
- ❑ Data providers must see the benefit to participate in the network
- ❑ Data flagging and data cleaning tools are key to support the identification of data inconsistencies
- ❑ Stable and long term funding is necessary to ensure development and the persistency of open and free data networks (persistent repositories are critical; funding mechanisms need to be improved !)

Putting the pieces together



the pieces put together



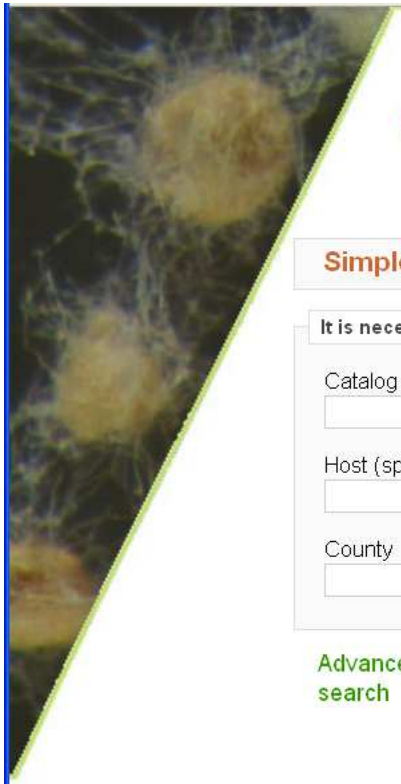
global catalogue

distribution maps

indicators reports

datacleaning reports


CRB
o projeto
temas associados
catálogo virtual
IICAMP
Embrapa
UNICAMP
Simple search
1 | Collections
| Next
Use the options below to select the collections to be searched
Uncheck all
Collection type
all
Collection location
Brazil
all
Quality level ???
BGB - Banco de Germoplasma de Bacillus spp. para controle biológico
BR - Coleção de Culturas de Bactérias Diazotróficas
CBMAI - Coleção Brasileira de Microrganismos de Ambiente e Indústria
CCFF - Coleção de Culturas de Fungos Filamentosos do Instituto Oswaldo Cruz
CCGB - Coleção de Culturas de Bacillus e Gêneros Correlatos
CFAF - Coleção de Culturas de Fitopatógenos e Agentes de Controle Biológico de Fitopatógenos
CG - Coleção de Culturas de Fungos Entomopatogênicos
CLIOC - Coleção de Leishmania do Instituto Oswaldo Cruz
IBSBF - Coleção de Culturas de Fitobactérias do Instituto Biológico
INCQS - Coleção de Microrganismos de Referência do Instituto Nacional de Controle de Qualidade em Saúde
SICol
português | help
Advanced search
Indicators
dataCleaning
Ministério da Ciência e Tecnologia, MCT
Centro de Referência em Informação Ambiental, CRIA



CRB
o projeto
temas associados

catálogo virtual

Simple search 1 | Collections 2 | Filters | Next

It is necessary to fill out at least one filter to carry out a search

Catalog number

Scientific name
bacillus

Host (species) or Substrate

Strain applications

Strain properties

County

State or province

Country

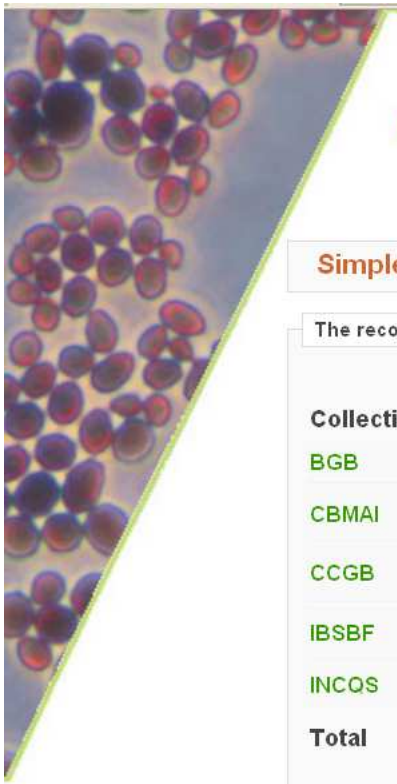
Advanced search

Indicators dataCleaning

Ministério da Ciência e Tecnologia, MCT
Centro de Referência em Informação Ambiental, CRIA

SICol

português | help



CRB
o projeto
temas associados

catálogo virtual

Simple search 1 | Collections 2 | Filters 3 | Results

The records found are presented per collection. Choose the content type, format, and desired output (text, maps, and Google maps).

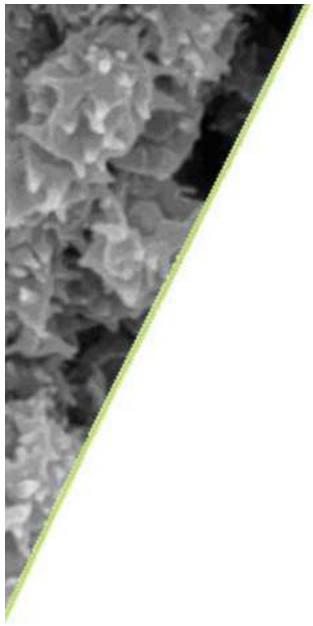
Collection	Records	Georeferenced records: At source	Automatic	Content	Format	Output
BGB	1989	0	0	Short	MS-Excel	see
CBMAI	67	0	3	Complete	XML	see
CCGB	837	0	71	Complete	HTML	see
IBSBF	25	0	1	Short	KML	see
INCQS	43	0	0	Short	HTML	see
Total	**2961**	**0**	**75**	Inventory	State or province	HTML	see

Automatic georeferencing is carried out for records from Brazil that do not have geographic coordinates, but have information on municipality. Records that have geographic coordinates blocked by curators are not georeferenced by the tool, respecting the decision that this is sensitive data. The system adopts the coordinates determined for the municipality by IBGE (Brazilian Institute of Geography and Statistics).

"Suspect" records are those that the geographic coordinate is not consistent with country, state, and/or municipality data according to the IBGE database. They are points that don't fall within the geographic space registered by the collection. The system does not evaluate whether the species is aquatic or terrestrial. Therefore if a record has "Brazil" as the country data and the coordinate falls within Brazil's

SICol

português | help



catálogo virtual

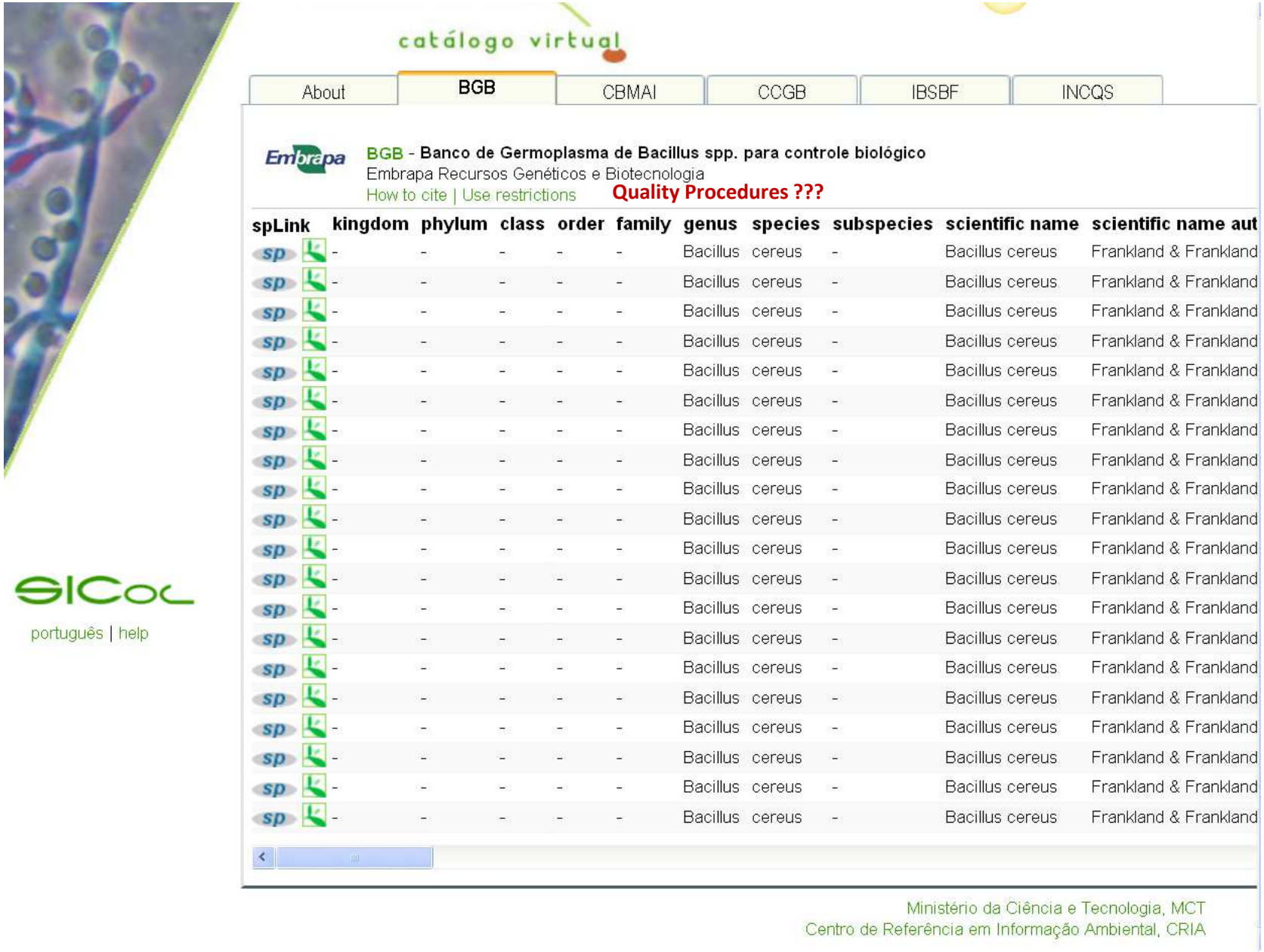
About | Inventory

Collections:
BGB, CBMAI, CCGB, IBSBF, INCQS

state or province	total
	958
AC	12
AL	10
AM	46
Amapá	4
Amazonas	20
AP	16
BA	101
CE	28
DF	241
ES	40
GO	247
Goiás	3
MA	26
Mato Grosso	1
Mato Grosso do Sul	3
MG	90
Minas Gerais	1
MS	95
MT	119
PA	105
Paraíba	1
PB	26
PE	58

SICol

português | help


catálogo virtual
About
BGB
CBMAI
CCGB
IBSBF
INCQS
Embrapa
BGB - Banco de Germoplasma de Bacillus spp. para controle biológico
Embrapa Recursos Genéticos e Biotecnologia
How to cite | Use restrictions
Quality Procedures ???
spLink kingdom phylum class order family genus species subspecies scientific name scientific name aut
sp - - - - - Bacillus cereus - Bacillus cereus Frankland & Frankland
SICol
português | help
Ministério da Ciência e Tecnologia, MCT
Centro de Referência em Informação Ambiental, CRIA

SICol

dataCleaning

português

This tool aims at helping curators in identifying possible errors and to standardize data. Records are not modified. The system just presents "suspect" records, recommending that they be checked by each author or curator. The tool is under constant development, so any suggestion is more then welcome.

Select a collection INCQS

collection: **INCQS**	
total number of records on-line	723
- without coordinates	723
- georeferenced	0
- access to georeferenced data denied	0
- in the sea	0
repeated records	
catalog number	0
all fields	0
collector's name and number	0
last update	
of the collection	02-02-2005
of dataCleaning	21-09-2010

geographic distribution of the specimens

this collection does not have georeferenced data

collection profile
dataCleaning statistics

taxonomic data	
inventory	scientific name - collector - types
family	not found
genus	not found
species	9 suspect records
subspecies	not found
author	not found
duplicate	not found

locality data	
inventory	country - state - municipality
name of the country/state	not found
outlier	not found
long/lat outside the world limit	not found
equal long/lat	not found
long or lat equal to zero	not found
long/lat in the sea (Brazil)	not found
municipality name (Brazil)	not found
coordinate unit analysis (Brazil)	not found

date collected	
collect before 1930	not found
last update previous to date collected	not found

suggestions for blank fields	
long/lat (Brazil)	not found
country/state name	not found
municipality name (Brazil)	not found

Scientific Name Search

Genus: *Corynebacterium* | Species: *dyphteriae* | subspecies: | in: All | search

português

Help | **Dictionaries** | Resources at CRIA | External Resources

DSMZ | CoL | Credits

DSMZ Bacteria (ver. July 2010)

1 Phonetically similar name found

Bacteria	Family
Corynebacterium diphtheriae (Kruse 1886) Lehmann and Neumann 1896 species (AL). Ref: IJSB 30:285 (AL). Risk group: 2. Strains: ATCC 27010.	Not assigned

0 sec

☐ Include synonyms ☑ Check systems at CRIA ☑ Check external systems Show 200 records

Memórias do Instituto Oswaldo Cruz
Fundação Oswaldo Cruz, Fiocruz
ISSN: 1678-8060
EISSN: 1678-8060
Vol. 98, No. 8, 2003, pp. 987-993

Memórias do Instituto Oswaldo Cruz, Vol. 98, No. 8, 2003, pp. 987-993

en

Diphtheria Remains a Threat to Health in the Developing World An Overview

Ana Luíza Mattos-Guaraldi; Lílian Oliveira Moreira; Paulo Vieira Damasco & Raphael Hirata Júnior

Abstract

Changes in the epidemiology of diphtheria are occurring worldwide. A large proportion of adults in many industrialized and developing countries are now susceptible to diphtheria. Vaccine-induced immunity wanes over time unless periodic booster is given or exposure to toxigenic *Corynebacterium diphtheriae* sp occurs. Immunity gap in adults coupled with large numbers of susceptible children creates the potential for new extensive epidemics. Epidemic emergencies may not be long in coming in countries experiencing rapid industrialization or undergoing sociopolitical instability where many of the factors thought to be important in producing epidemic such as mass population movements and difficult hygienic and economic conditions are present. The continuous circulation of toxigenic *C. diphtheriae* emphasizes the need to be aware of epidemiological features, clinical signs, and symptoms of diphtheria in vaccine era so that cases can be promptly diagnosed and treated, and further public health measures can be taken to contain this serious disease. This overview focused on worldwide data obtained from diphtheria with particular emphasis to main factors leading to recent epidemics, new clinical forms of C. diphtheriae infections, expression of virulence factors, other than toxin production, control strategies, and laboratory diagnosis procedures.

Keywords

adherence* - *Corynebacterium diphtheriae* sp - *diphtheria in adults* - *epidemics* - *laboratory diagnosis

Scientific Name Search

Genus: *Corynebacterium* | Species: *diphtheriae* | subspecies: | Kingdom/Domain: Bacteria | search

português

Help | Dictionaries | Resources at CRIA | **External Resources**

LPSN | BiotaNeotropica

J.P. Euzéby: List of Prokaryotic Names with Standing in Nomenclature.

J.P Euzéby: List of Prokaryotic Names with Standing in Nomenclature

See matches

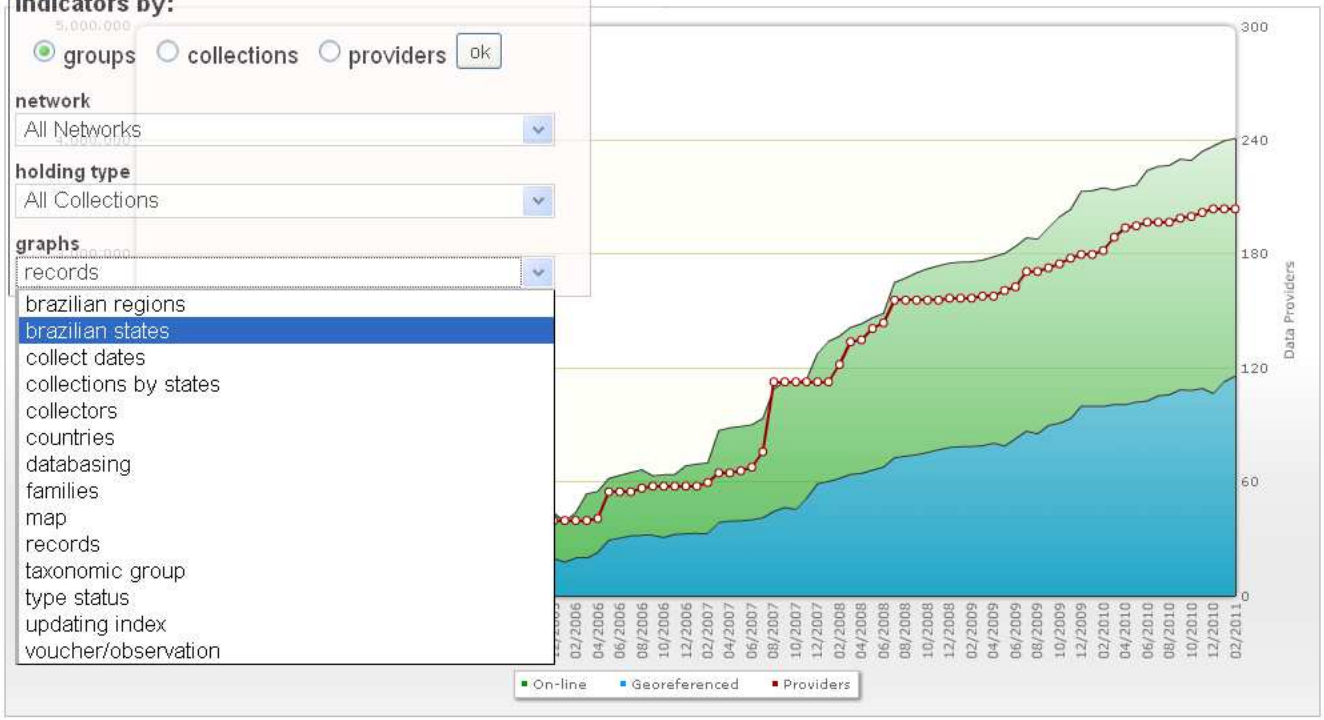
☐ Include synonyms ☑ Check systems at CRIA ☑ Check external systems Show 200 records

Indicators

All indicators presented are based on on-line data only. They are dynamic or daily reports presented as charts or graphs. The indicator reflects only the analysis of data that is available on-line so therefore may not reflect the reality of each collection, especially those with a low percentage of digitized data.**Click here to view the table data contribution x dependency per state.**

See other options for indicators here

All Networks - All Collections - records



Indicators

See other options for indicators here

SICol Network - All Collections - brazilian states

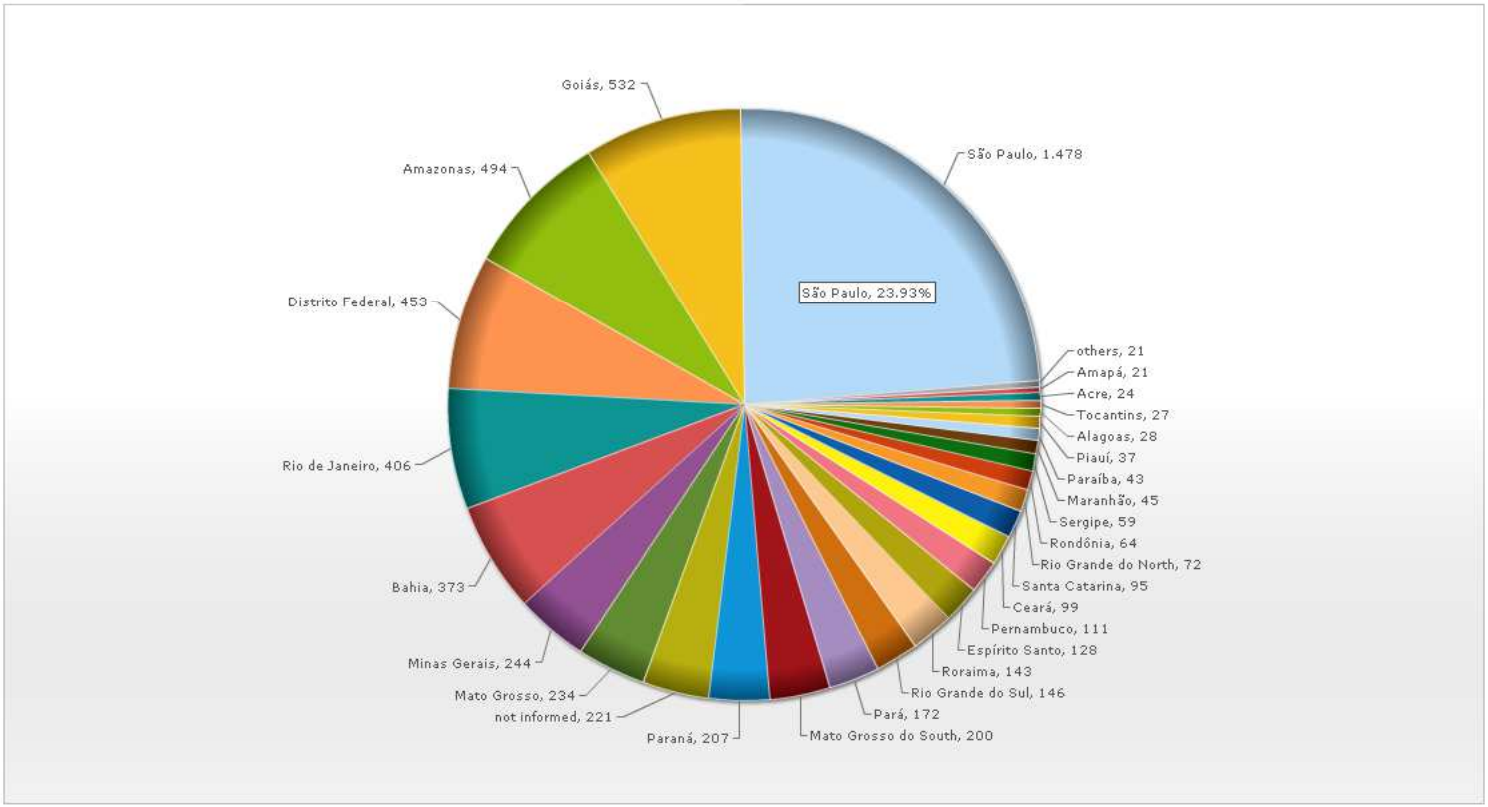


Chart with the percentage of records for each Brazilian State. As holdings are digitized and made available in the network, this indicator will show the number of records per state, possibly showing geographic information gaps.

Updated on 25/01/11 04:37

See other options for indicators here

SICol Network - All Collections - databasing

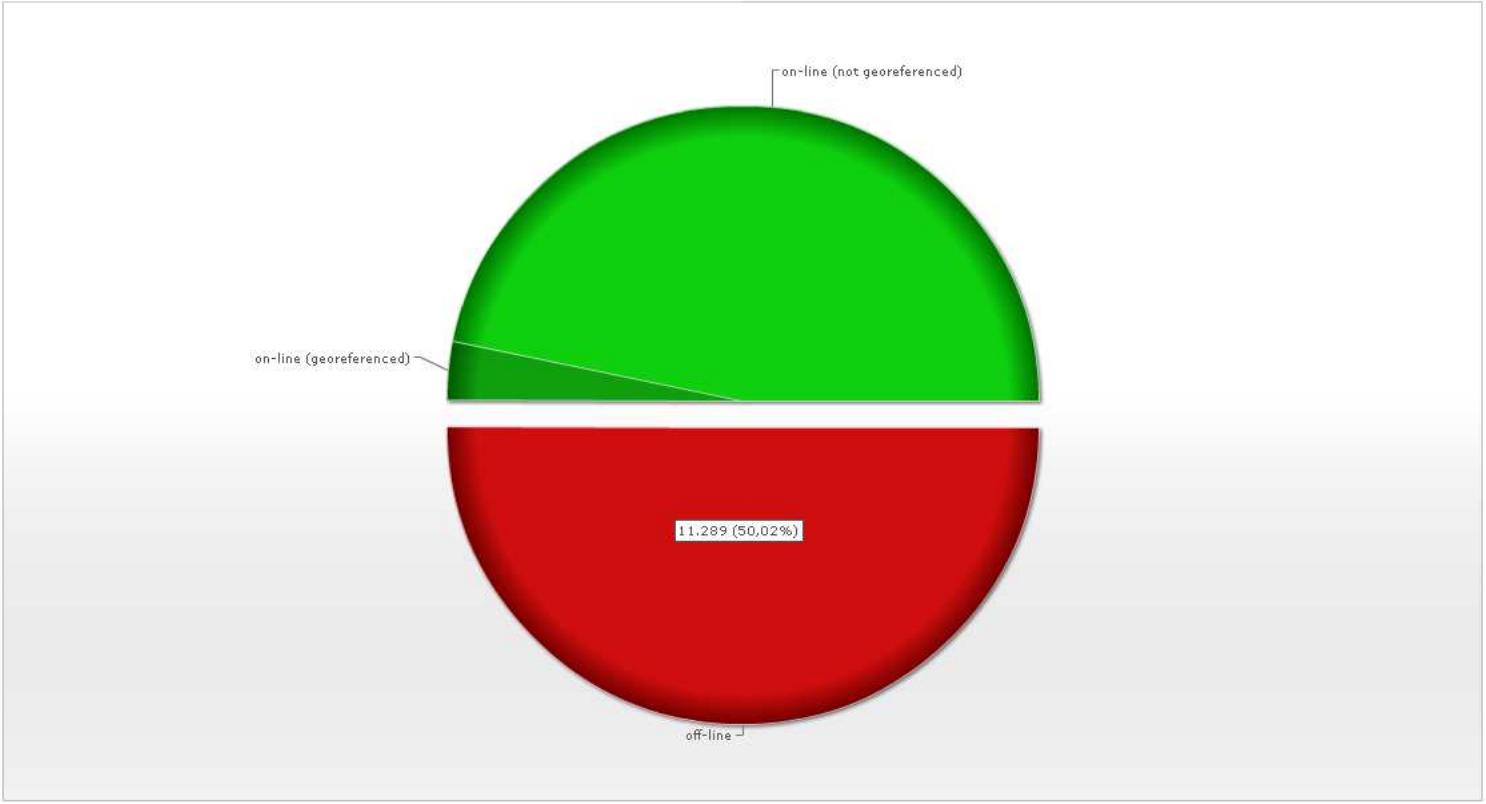


Chart with the percentages of on-line and georeferenced records compared to total records. This indicator aims to show the work still required to digitize and georeference collections' holdings.

Updated on 11/02/11 04:21

Making things possible

- Setting big goals, but…
  - Step by step approach
  - Problem driven approach

- Make it simple or at least doable

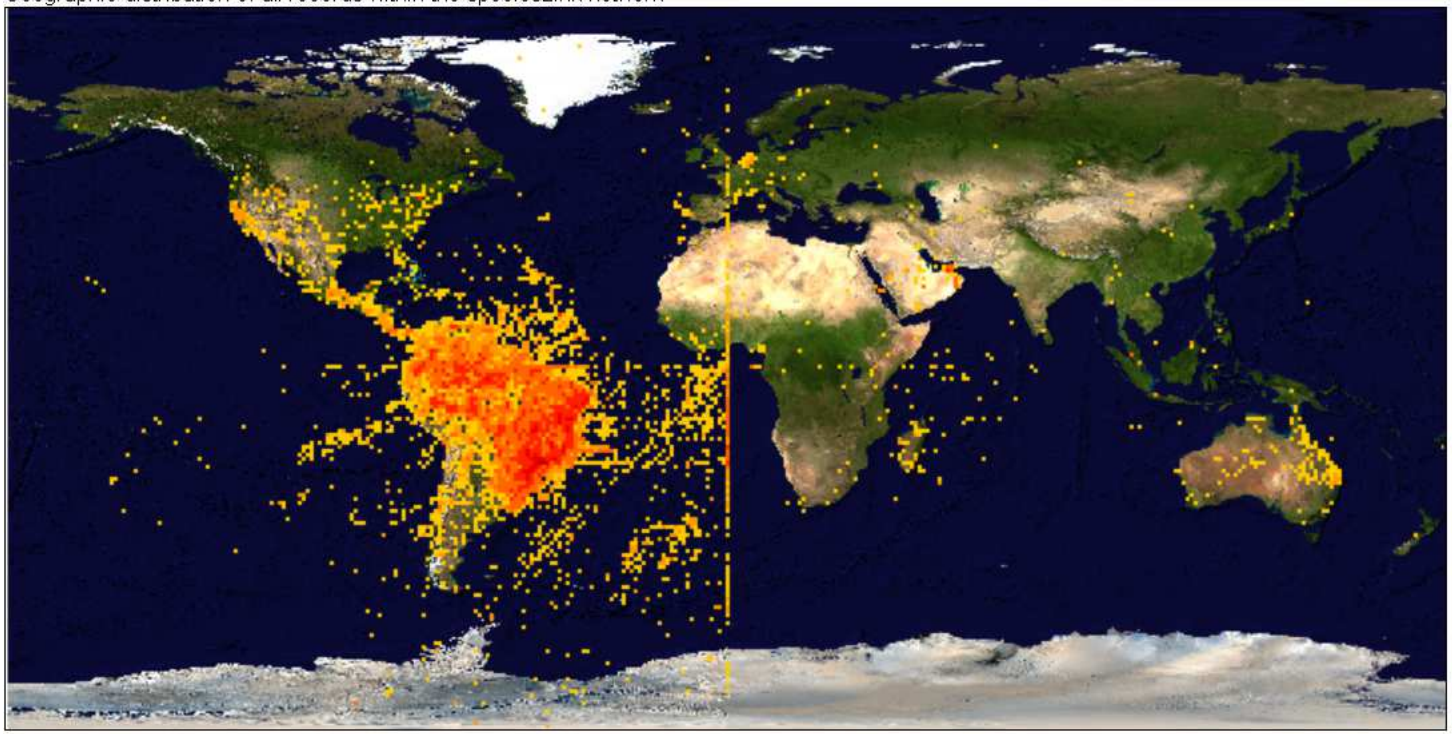
Sponsors & Funders

International Partners

National Partners

Thank you

Geographic distribution of all records within the speciesLink network



Dora Ann Lange Canhos
dora@cria.org.br