Challenges in building an infrastructure for all of biology

The importance of local infrastructures

Dora Ann Lange Canhos (dora@cria.org.br)

CRIA – Reference Center on Environmental Information

Some challenges …

My bias: species & specimen data, Brazil, S&T Policy

- Focus: content, users, & output (doable, step-by-step approach)
- Useful & Usable
- Known quality
- Scale: local to global
- Challenges are both local and global
- Integrate and serve data
- Technology: standards, protocols, bandwidth, processing, storage …

Some challenges …

- Show past, present, future, and trends
- Dynamic: open to new questions and insights
- Infrastructure(s) to help tackle a broad range of scientific and societal problems
- Integrate different sub-disciplines of biology
- (but also) promote the integration of specialists from different fields of knowledge, different countries, cultures, questions, insights, …

Some challenges …

- Governance: from centralized systems to networks of networks
- Project vs long-term
- Sustainability (long-term)
- Platform to share data, information, knowledge, and to interact and build new knowledge
- … and user/machine friendly …

E-infrastructures are not only important to support research, but are becoming vital to conduct research and to develop and monitor policies and strategies by diverse publics

The Levels of Complexity of the Biosphere

Biosphere - The world we live in

Ecosystem - The set of communities of all domains of life that interacting with one another and the abiotic environment to form a unit (e.g., freshwater ecosystems, taigas)

Community - Interacting populations of organisms (e.g., coral reefs, montane forest)

Population - All individuals of a species or phylotype within a community (e.g., trees of a given species within a single forest, the fishes of a given species in a single coral reef)

Organism - An single individual (e.g., a lizard, a tree, a bacterium)

Organ system - A specialized functional system of an organism (e.g., digestive, nervous)

Organ - A set of tissues that function as a unit (e.g., heart, brain, kidney)

Tissue - A set of interacting cells (e.g., epithelia, muscle)

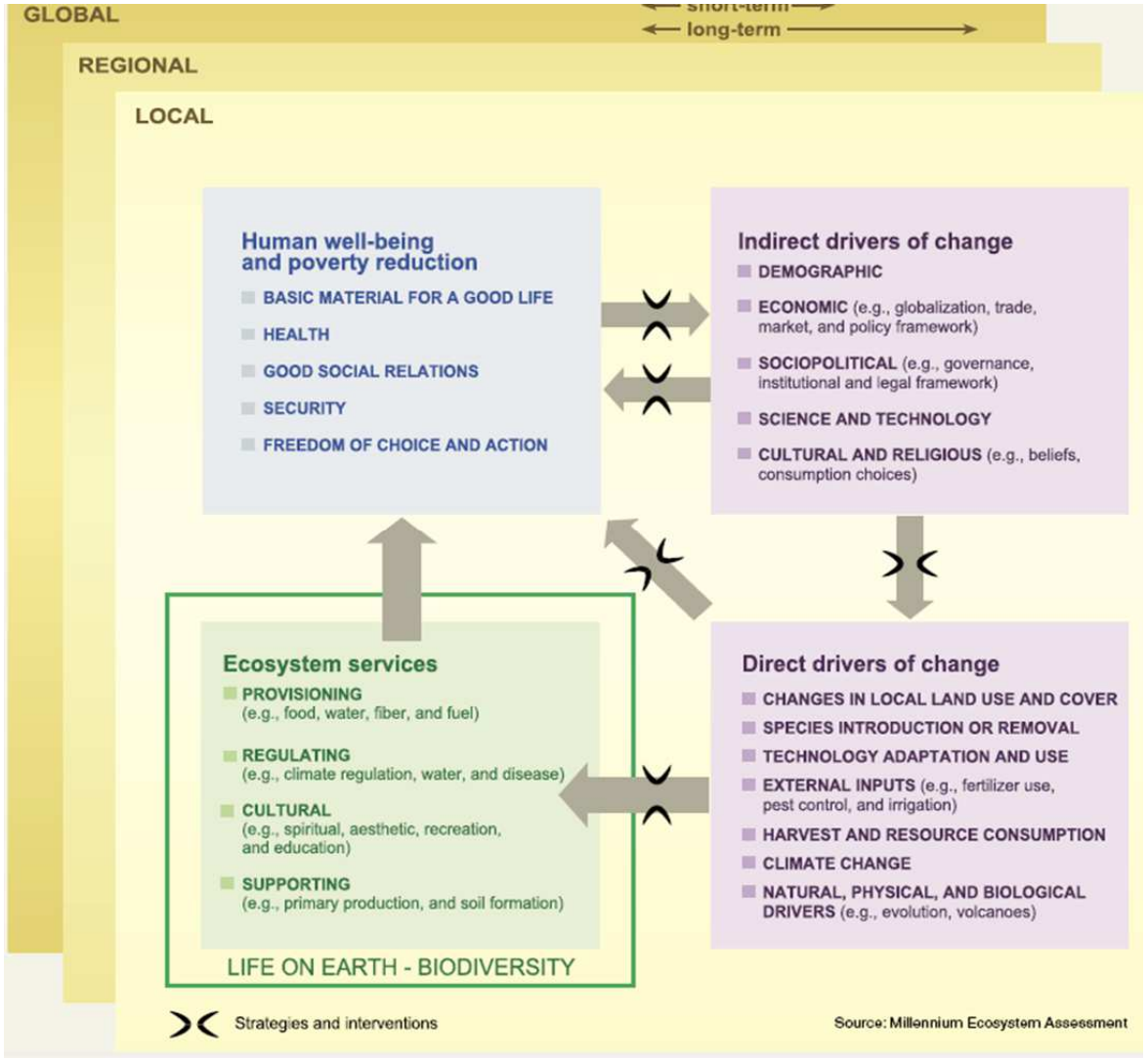
Cell - The functional unit of all living organisms (e.g., red blood cell, neuron, bacterium)

Organelle - A specialized subunit within cell (e.g., mitochondria, chloroplast)

Molecule - biochemical constituents of cells (e.g., a protein, a nucleic acid)

SOURCE: Committee on a New Biology for the 21st Century.

Millennium Ecosystem Assessment (2005)



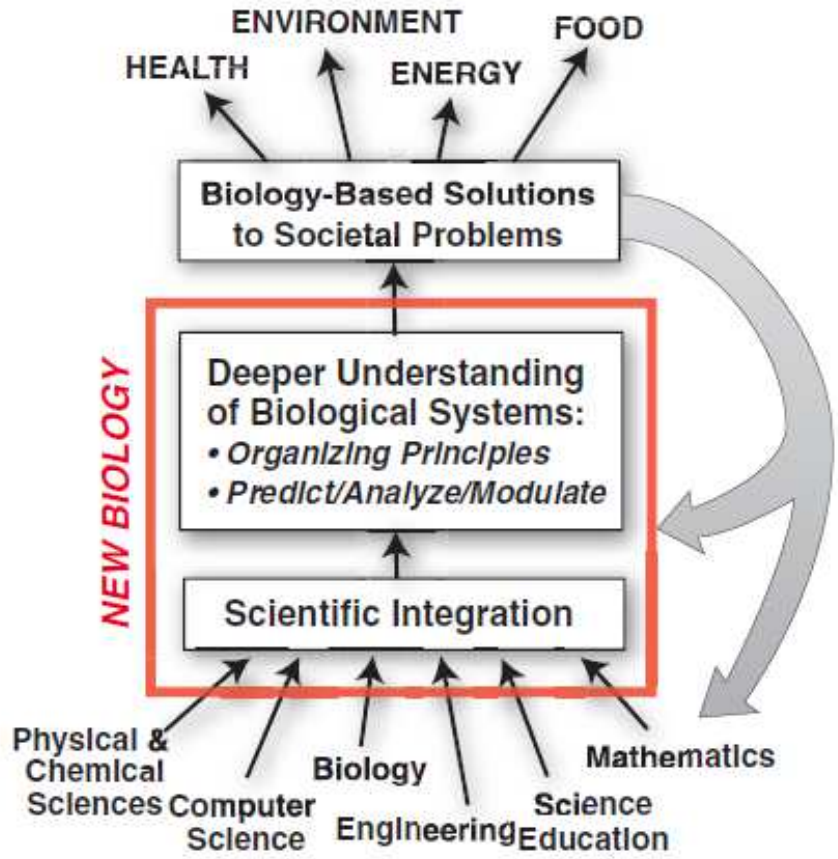
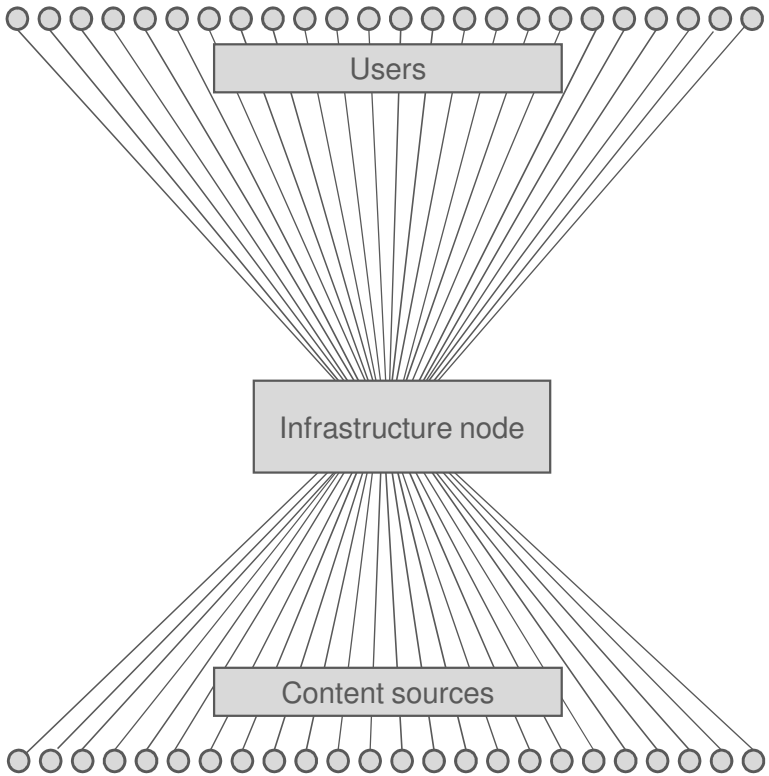


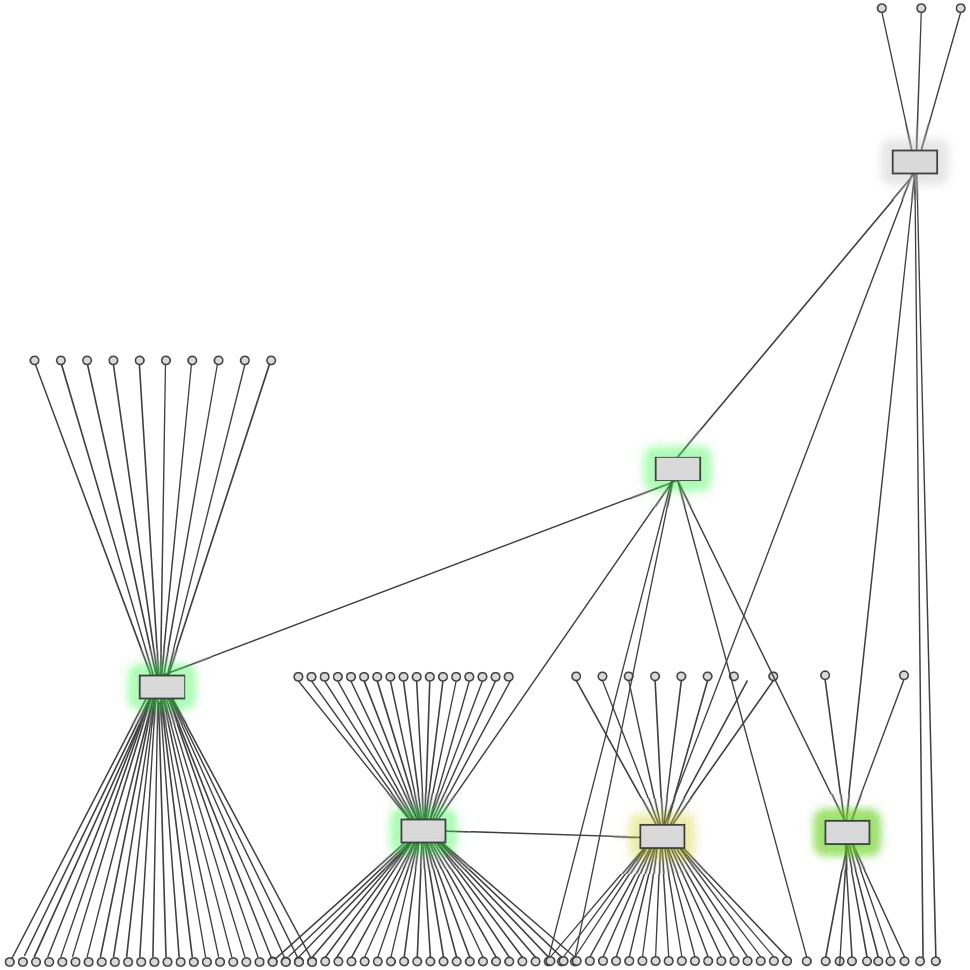
FIGURE 2.1 What is the New Biology?
SOURCE: Committee on a New Biology for the 21st Century.

Unifying Biology

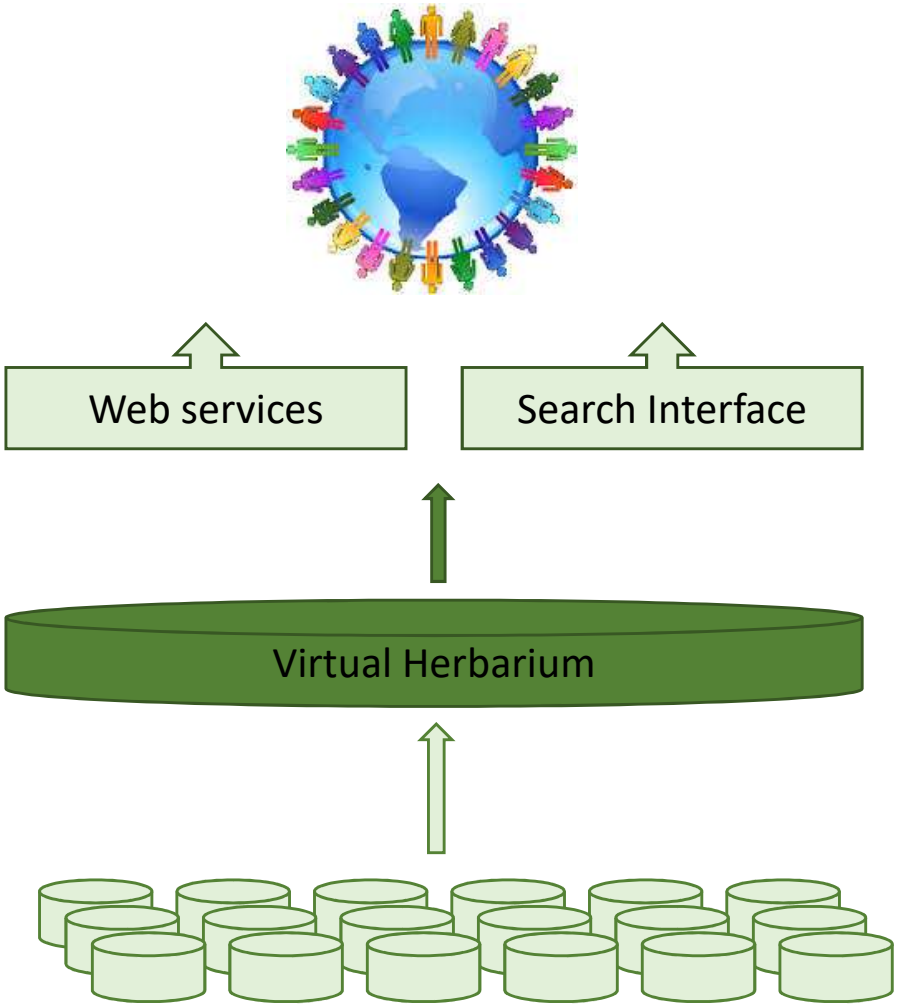
- Integrating different subdisciplines
- Promoting the integration of specialists from different fields of knowledge
- Targeting at both local and global problems such as sustainable development

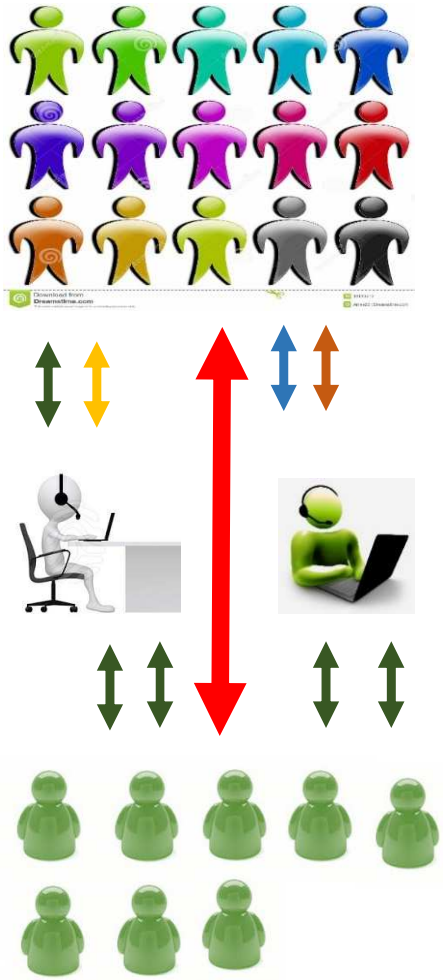
e-infrastructure networks

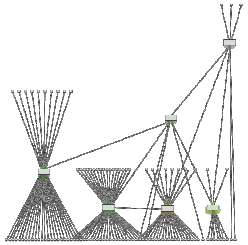




The importance of local e-Infrastructures







- species occurrence data
- Images (vouchers, live plants, pollen)
- maps, graphics, synthesis, catalogues (dynamic)

- data
- images
- maps
- ecological niche modeling

Search Interface

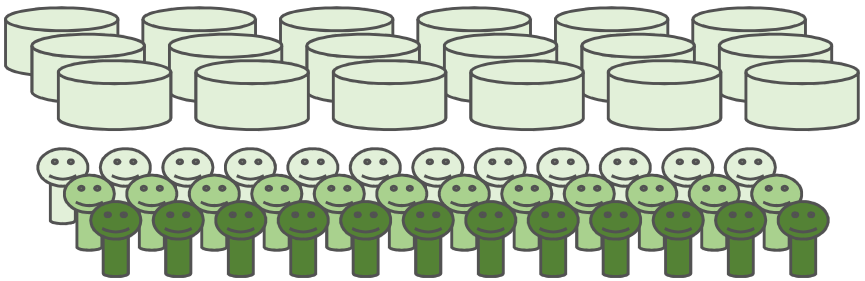
Web Services

INCT Virtual Herbarium of Flora and Fungi

Tools

- species occurrence data
- Images

- data cleaning
- automatic georeferencing
- indicators



- Cybertaxonomy
- Annotation system
- Lacunas
- BioGeo

Local infrastructures

- Focus on local needs
- Promote organization and dissemination of data
- Foster networking and collaboration

- Example: Brazil's Virtual Herbarium

Data Cleaning Reports

speciesLink — 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: ASE

collection: ASE

total number of records on-line	32796
- without coordinates	18
- georeferenced	32778
- access to georeferenced data denied	0
- in the sea	58
- blank catalognumber	1
smaller: 1 larger: 35763	[gap]

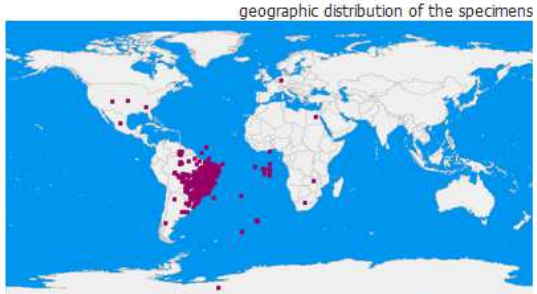
repeated records

catalog number	335
duplicate records	0
collector's name and number	1296

last update - error logs

of the collection: 12-09-2015 of dataCleaning: 13-09-2015

geographic distribution of the specimens



collection profile
dataCleaning statistics
geographic coordinates analysis

taxonomic data

inventory	scientific name - collector - identifier - types
family	not found
genus	not found
species	24 suspect records
subspecies	not found
author	142 suspect records
duplicate	1648 suspect records
other inconsistencies	not found

locality data

inventory	country - state - municipality
name of the country/state	610 suspect records
outlier	9 suspect records
long/lat outside the world limit	not found
equal long/lat	not found
long or lat equal to zero	12 suspect records
long/lat in the sea (Brazil)	27 suspect records
municipality name (Brazil)	3740 suspect records
coordinate unit analysis (Brazil)	11 suspect records
other inconsistencies	not found

date collected

collected before 1902	not found
last update previous to date collected	not found
identification year previous to date collected	not found

suggestions for blank fields

long/lat (Brazil)	11 suggestions
country/state name	22 suggestions
municipality name (Brazil)	211 suggestions

search
dataCleaning

email
Centro de Referência em Informação Ambiental, CRIA

Suspect Records & Inconsistencies

- Taxonomic data
- Dates
- Locality data
- Blank fields

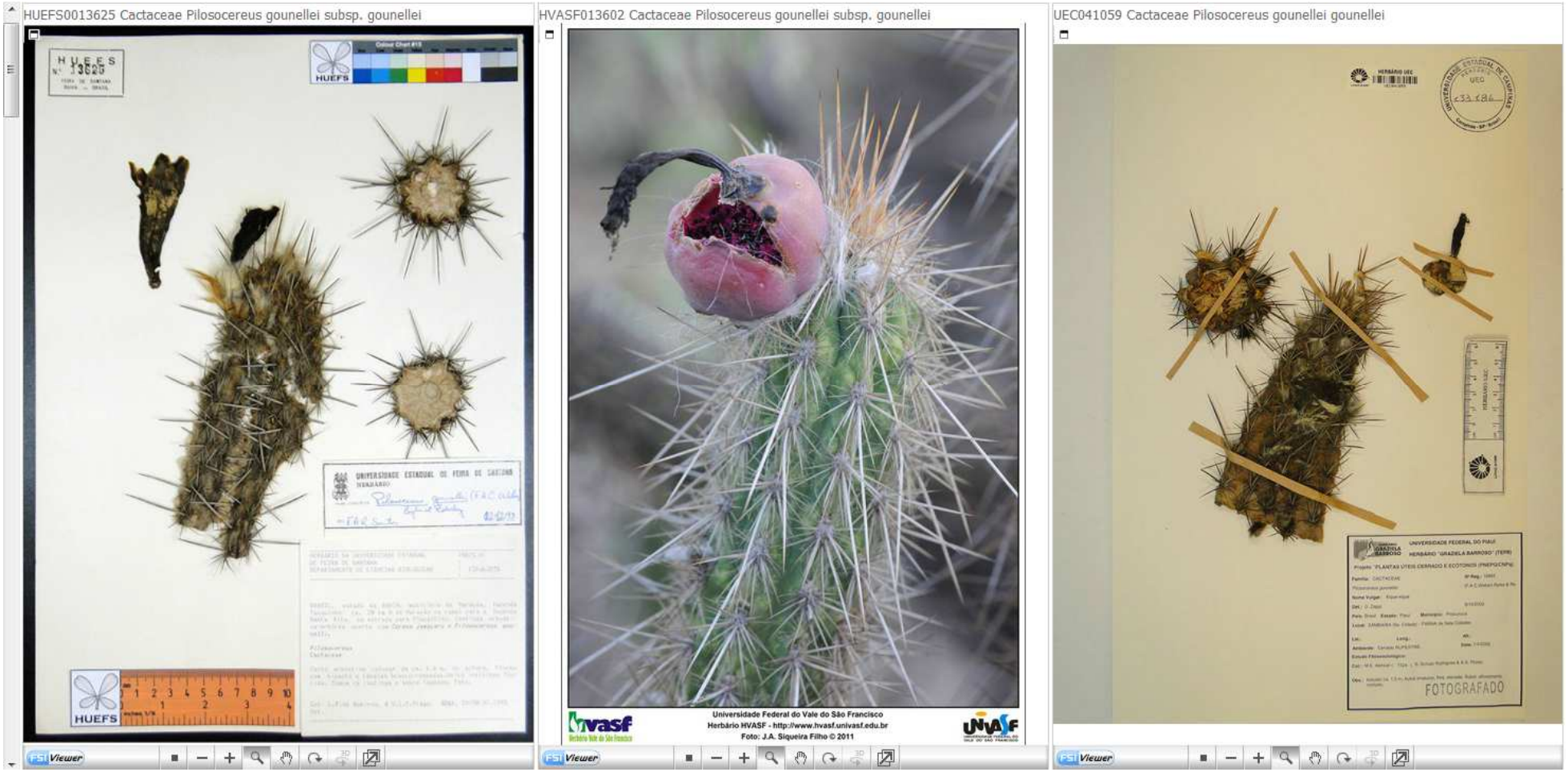
Data cleaning and duplicates

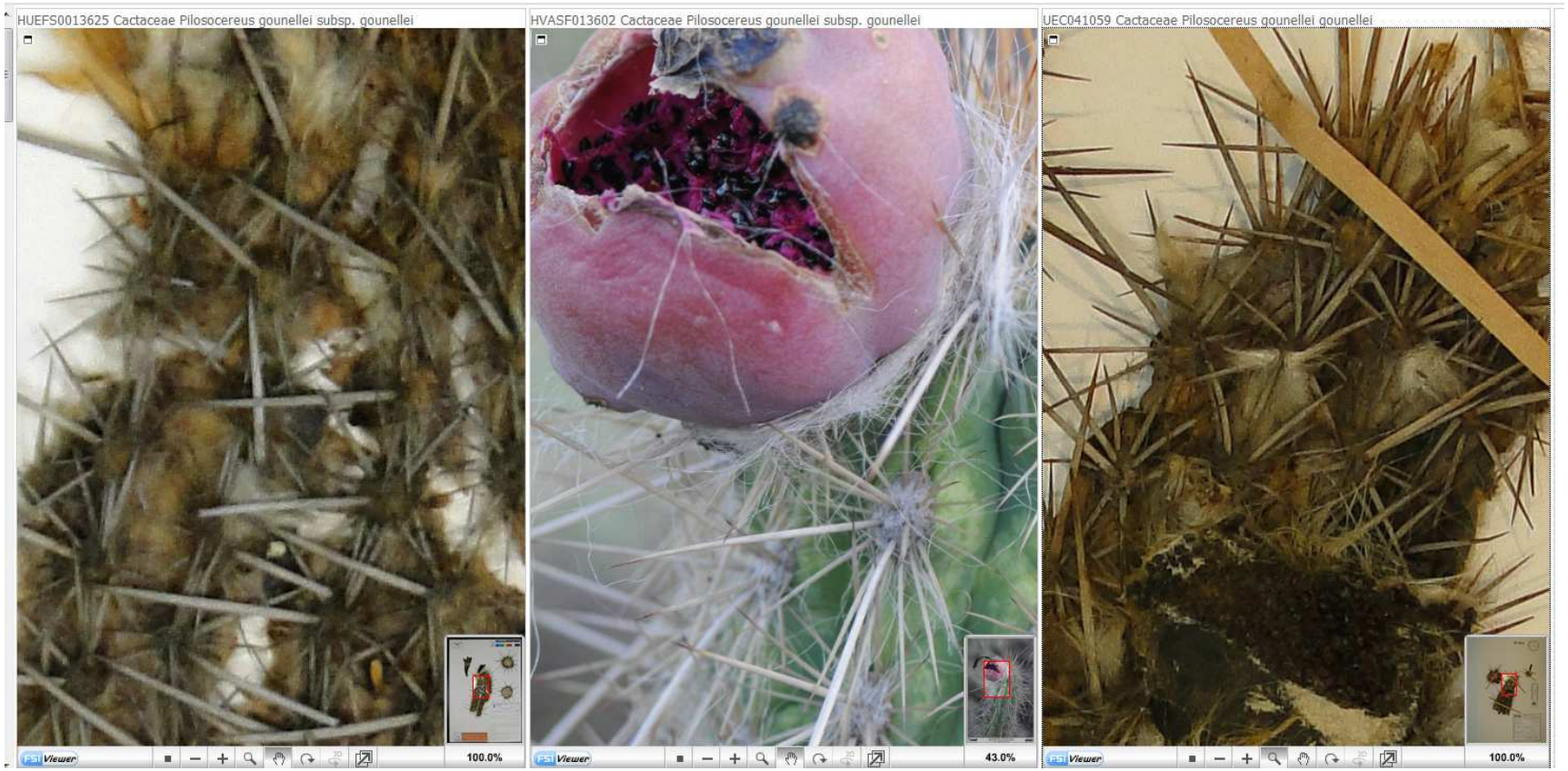
Filter: [no filter] [undetermined family] [undetermined genus] ☑ same collection date

collector	collector number	date collected	img	catalognumber	family	genus	species
Almeida, MN de	30	1982-01-23		ASE 2111	Solanaceae	sp *Solanum*	*stipulaceum*
Almeida, MN	30	1982-01-23		JPB 58260	Solanaceae	sp *Solanum*	*asperum*

identified by	date identified	stateprovince	county
Agra, MF	2012-04	Sergipe	São Cristóvão
Viana, G	1988-10-31	Sergipe	São Cristovão

Cybertaxonomy

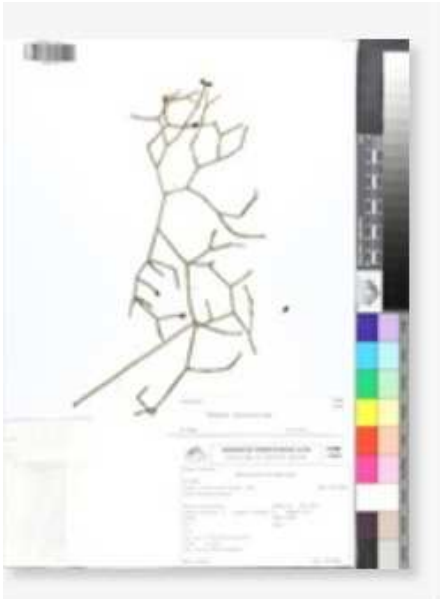



HUEFS0013625 Cactaceae Pilosocereus gounellei subsp. gounellei
100.0%
HVASF013602 Cactaceae Pilosocereus gounellei subsp. gounellei
43.0%
UEC041059 Cactaceae Pilosocereus gounellei gounellei
100.0%

Annotations

Herbário Dr. Roberto Miguel Klein - FURB
Universidade Regional de Blumenau, FURB

PLANTAE DICOTYLEDON CACTACEAE
Rhipsalis cereuscula Haw.. Det: Zappi, DC **13/12/2012**
FURB 11879 Collect: Dreveck, S; Carneiro, FE 386 **18/11/2008**.
Loc: Linha Santa Terezinha, Herval d'Oeste, Santa Catarina, Brasil Barcode: **FURB01187.**
Coord orig: [*lat:* -27.135 *long:* -51.435 WGS84] Altitude: 742m.
Notes: Florística IFFSC 2100(Epífita); epífita;; Floresta Estacional Decidual; Estágio Avançado
Basis of record: preserved specimen

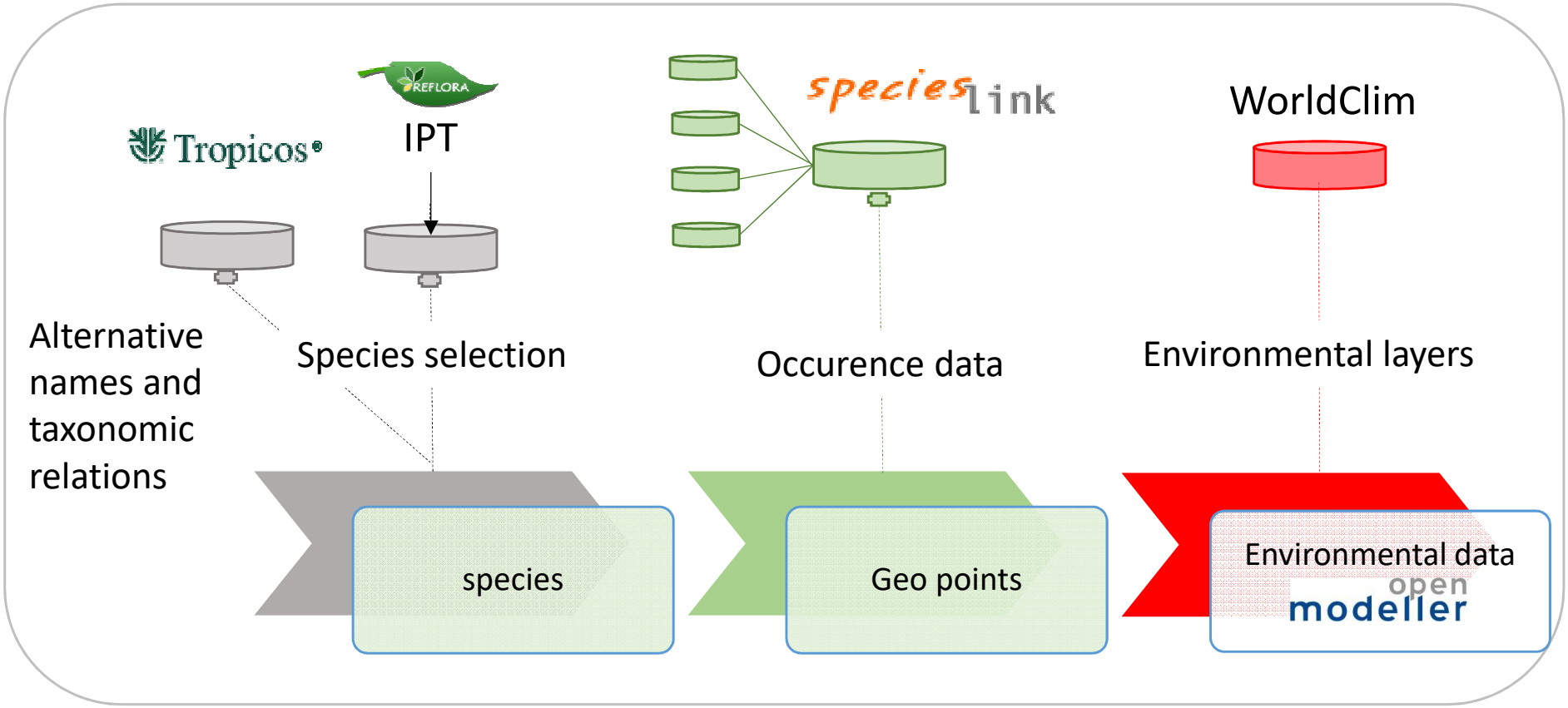


reference
CACTACEAE
Rhipsalis baccifera (J.S.Muell.) Stearn. Det: S. Dreveck **18/11/2008** Barcode: **FURB01187. FURB 11879** Collect: S. Dreveck 386 Loc: Linha Santa Terezinha Herval d'Oeste Santa Catarina Brasil **18/11/2008**. [*lat:* -27.135 *long:* -51.435 WGS84]

comment
sciname
Olá! Com certeza esse espécime não é R. baccifera, poderia enviar uma imagem em alta resolucao para que eu possa identificar corretamente? Muito obrigada, Zappi

Daniela Zappi
13-Dec-2012 00:16

Plataform for *e-Science: BioGeo*



BioGeo (ecological niche model)

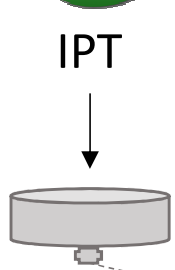
http://biogeo.inct.florabrasil.net

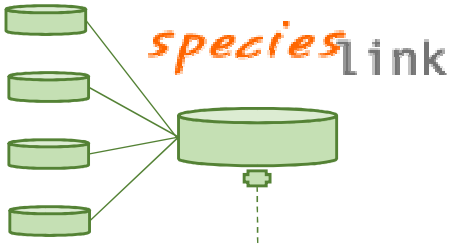
Accepted names, synonyms, states of occurrence

Status and origin of data states of occurrence

Portaria 443 (redlist)

IPT





Filter accepted names and synonyms (phonetic)

Lacunas database

IUBS

Lacunas: September 2012 – August 2015

Species with no data records: phonetic search including synonyms, with or without geographic coordinates

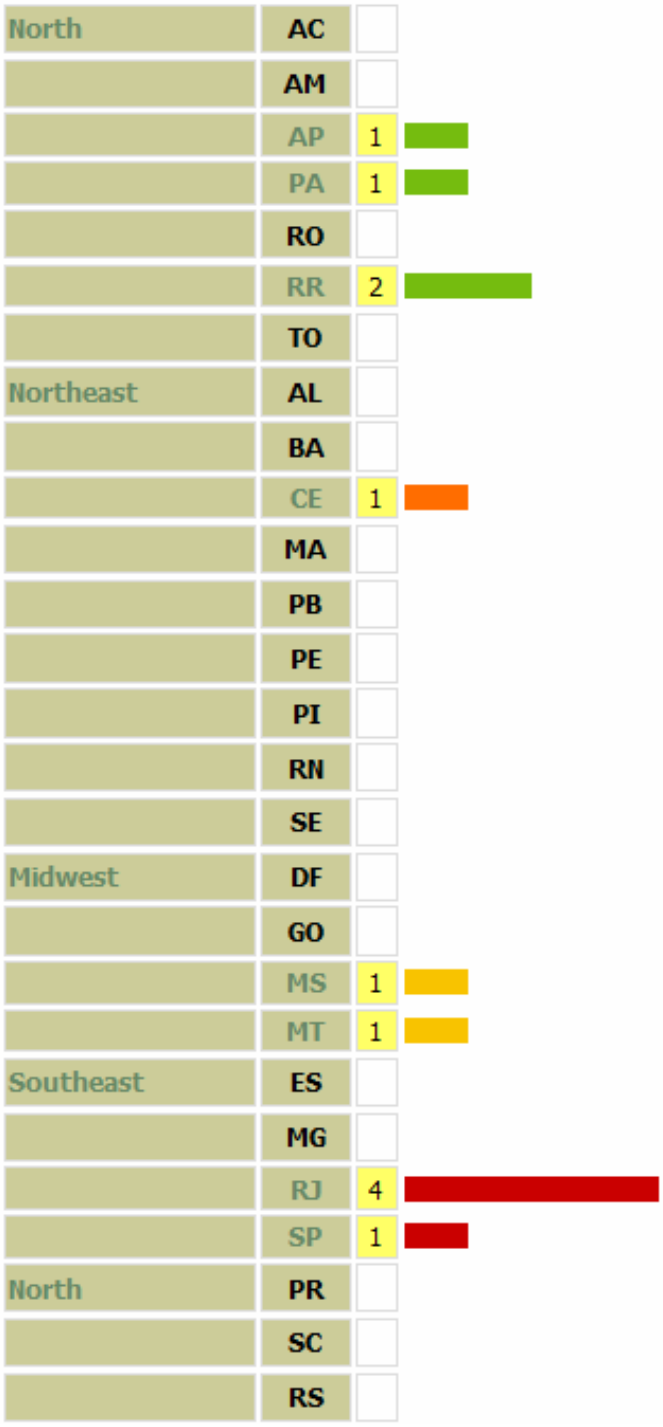
Taxonomic Groups	Brazil's Flora	Sep/2012		Aug/2015	
	No. Species	Species	%	Species	%
Algae	4.746	2.555	54%	2.424	51%
Angiosperms	32.824	2.429	7%	1.578	5%
Bryophytes	1.524	324	21%	196	13%
Fungi	5.710	2.925	51%	2.453	43%
Gymnosperms	30	0	0%	1	3%
Ferns & Lycophytes	1.253	121	10%	62	5%
Total	46.087	8.354	18%	6.714	15%

Lacunas: September 2012 – August 2015

Species > 20 records: phonetic search including synonyms, with consistent & distinct geographic coordinates (origin)

Taxonomic Groups	Brazil's Flora	September 2012		August 2015	
	No. Species	Species	%	Species	%
Algae	4.746	46	1%	103	2%
Angiosperms	32.824	5.587	17%	8.250	25%
Bryophytes	1.524	158	10%	246	16%
Fungi	5.710	13	0,2%	21	0,4%
Gymnosperms	30	4	13%	5	17%
Ferns & Lycophytes	1.253	274	22%	473	38%
Total	46.087	6.082	13%	9.098	20%

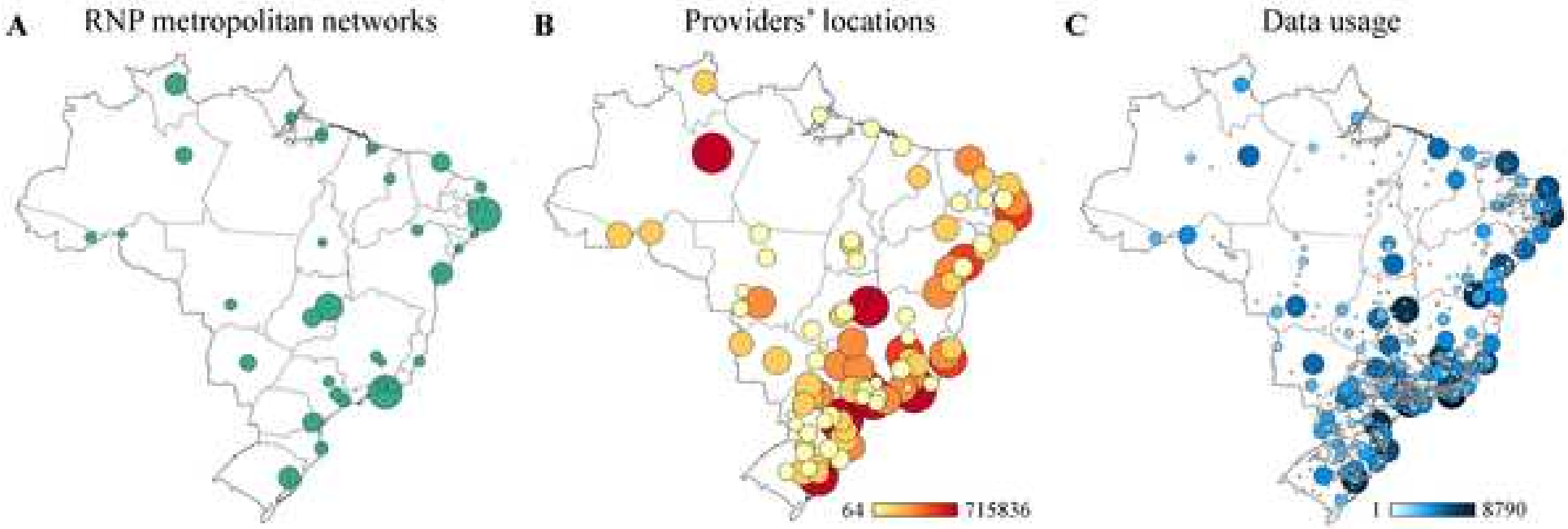
Geographic Gaps - Monimiaceae



- *Hennecartia omphalandra* J.Poiss.
- *Mollinedia calodonta* Perkins
- *Mollinedia myriantha* Perkins
- *Mollinedia sphaerantha* Perkins

Research infrastructure and biodiversity data usage in Brazil.

1.4 million records/day



Canhos DAL, Sousa-Baena MS, de Souza S, Maia LC, Stehmann JR, et al. (2015) The Importance of Biodiversity E-infrastructures for Megadiverse Countries. PLoS Biol 13(7): e1002204. doi:10.1371/journal.pbio.1002204
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002204

PLOS | BIOLOGY

IUBS

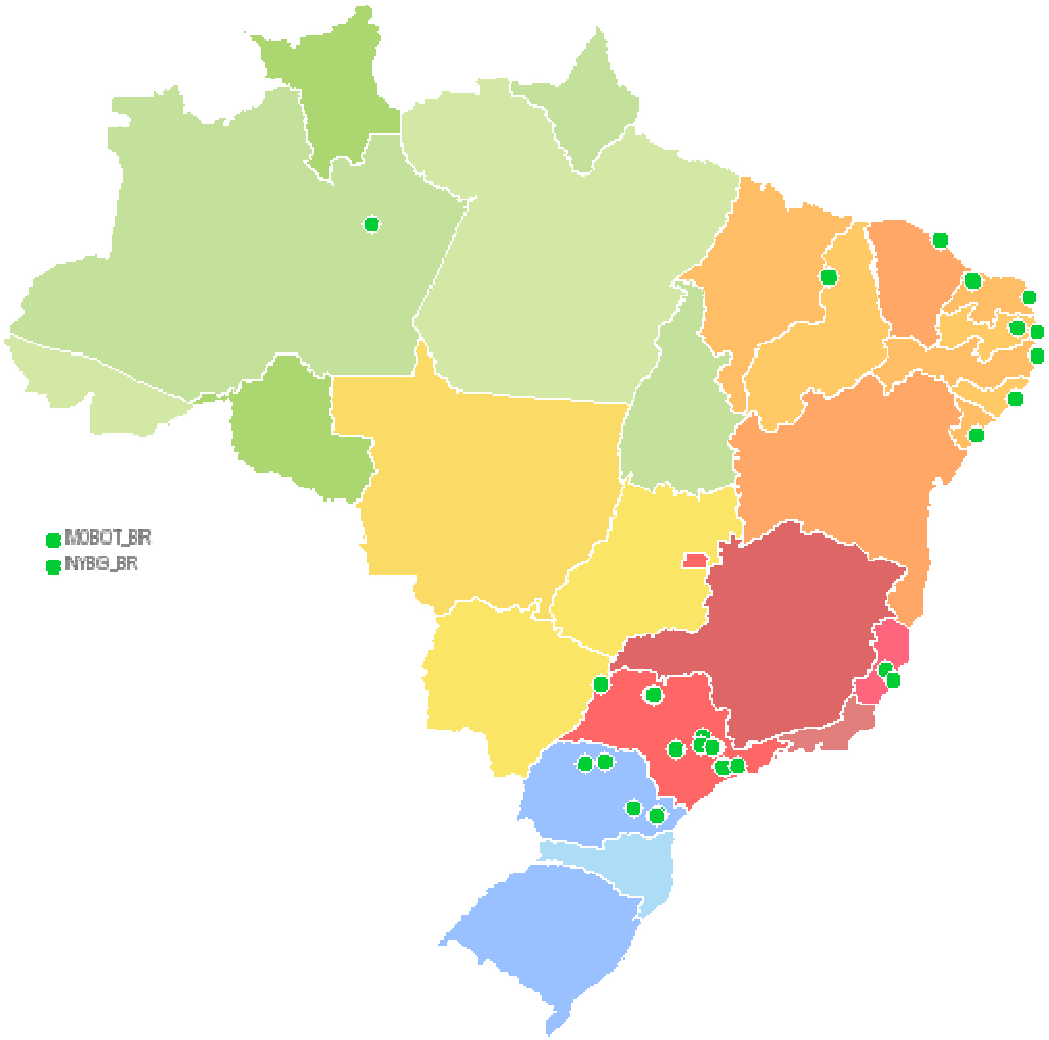
Outcomes from sharing data

- greater institutional recognition
- greater involvement with graduate courses
- increased number of visits to the herbaria
- increase of the holdings
- increase of grants

OPEN AND COLLABORATIVE SCIENCE IN DEVELOPMENT NETWORK

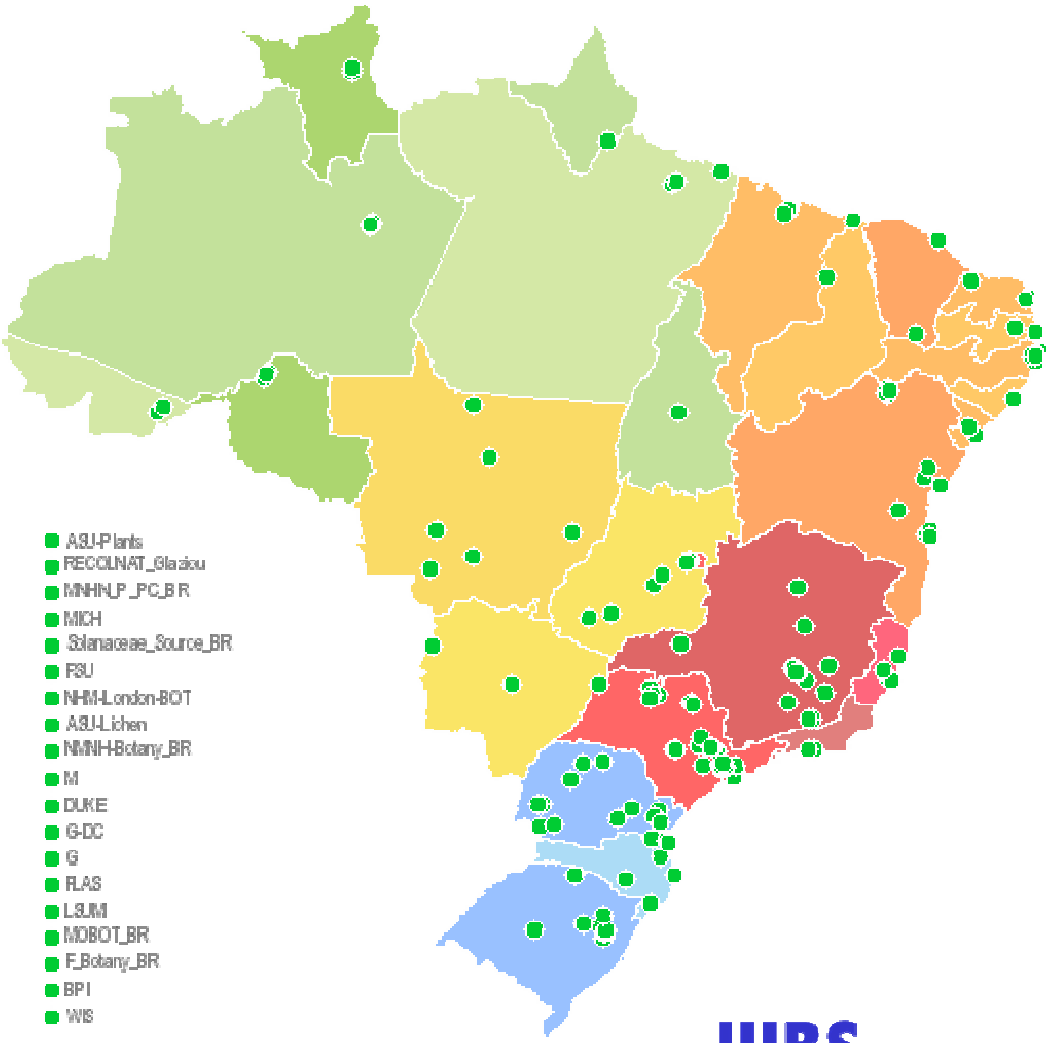
Brazil's Virtual Herbarium

48 datasets – 2 from abroad



2008

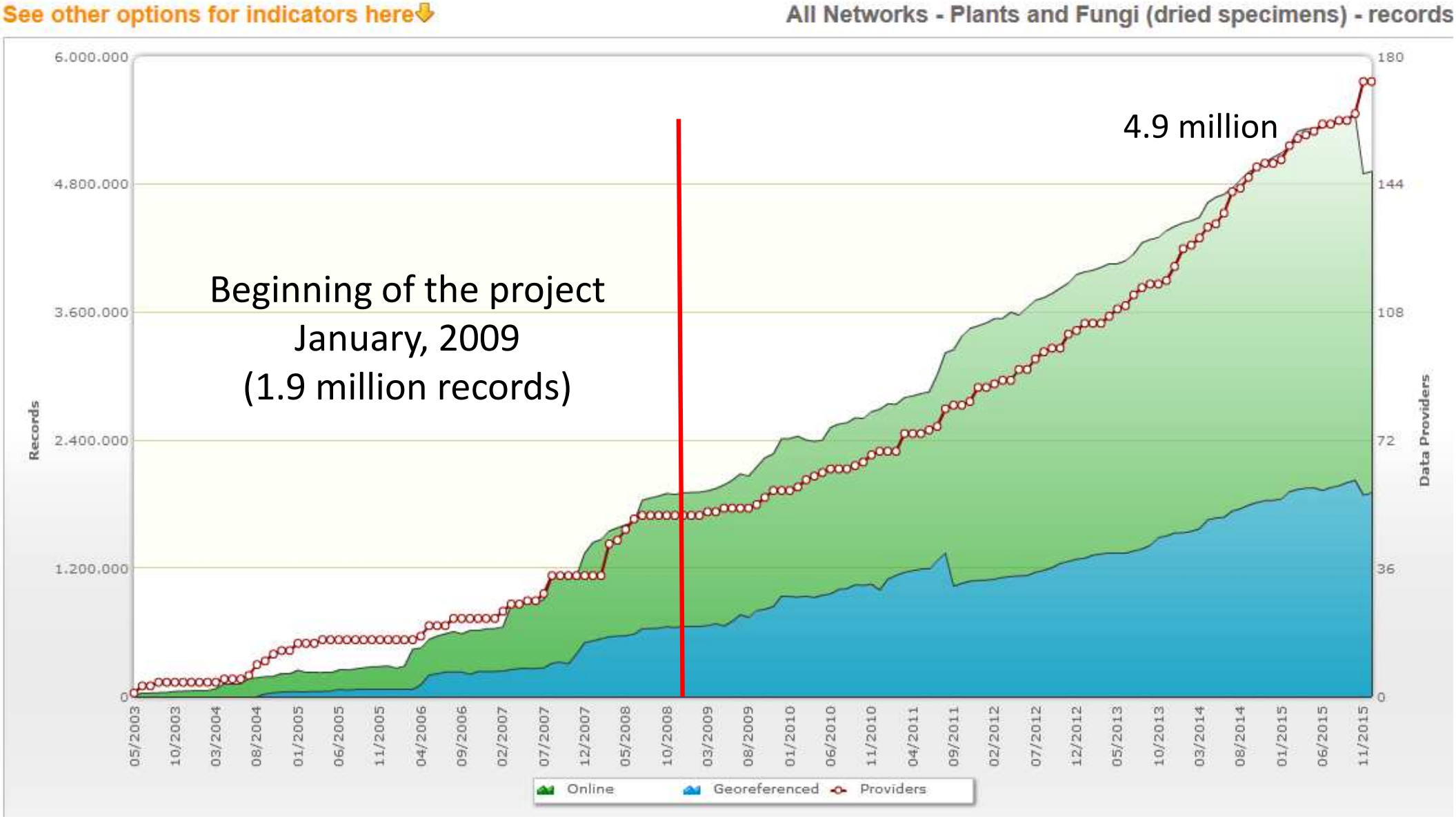
174 datasets – 24 from abroad



2015

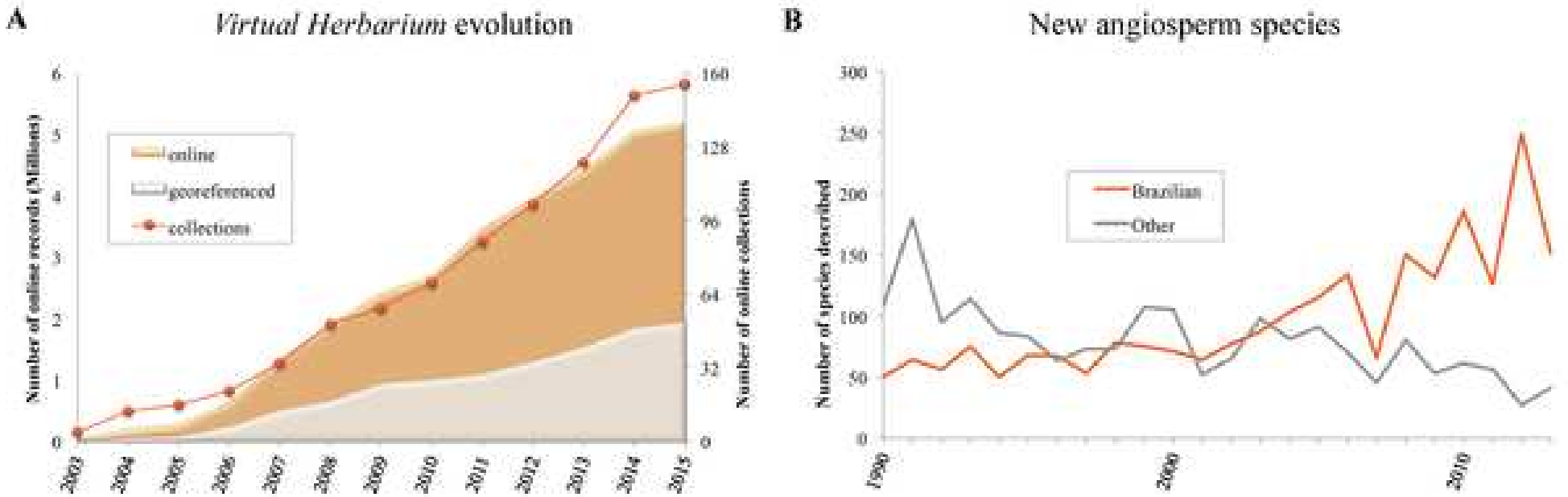
Indicators

See other options for indicators here



> 79K distinct species > 1 million images

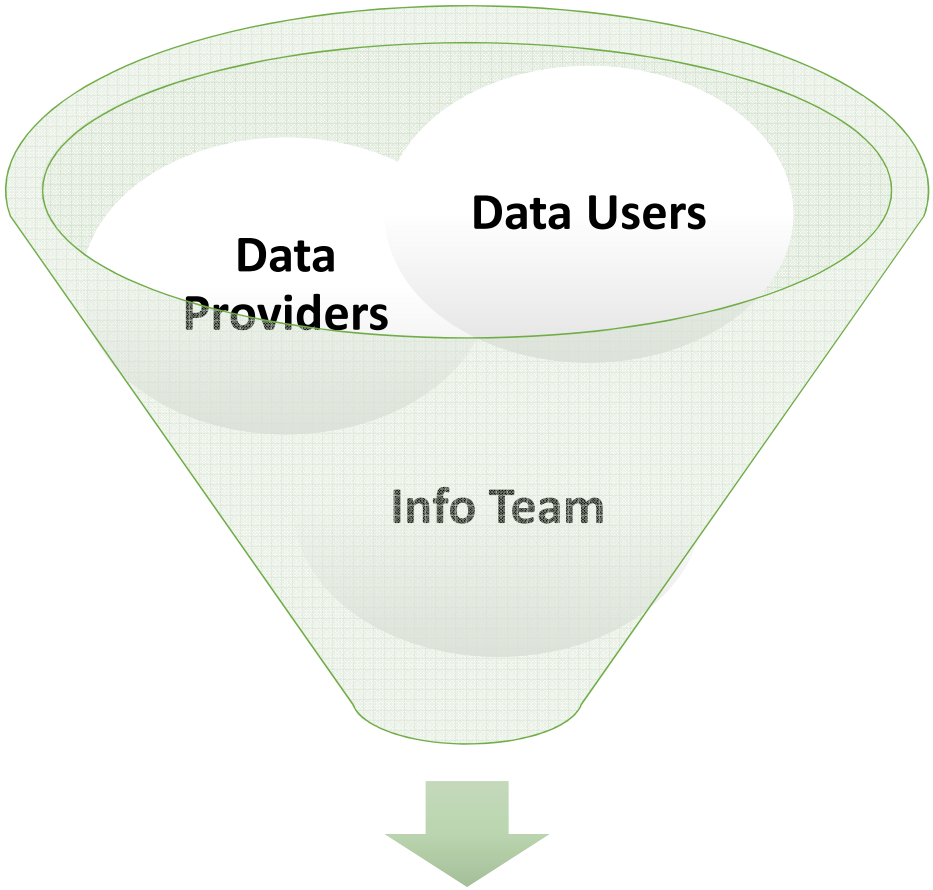
The evolution of Brazilian flora primary data available online and species described by Brazilian specialists



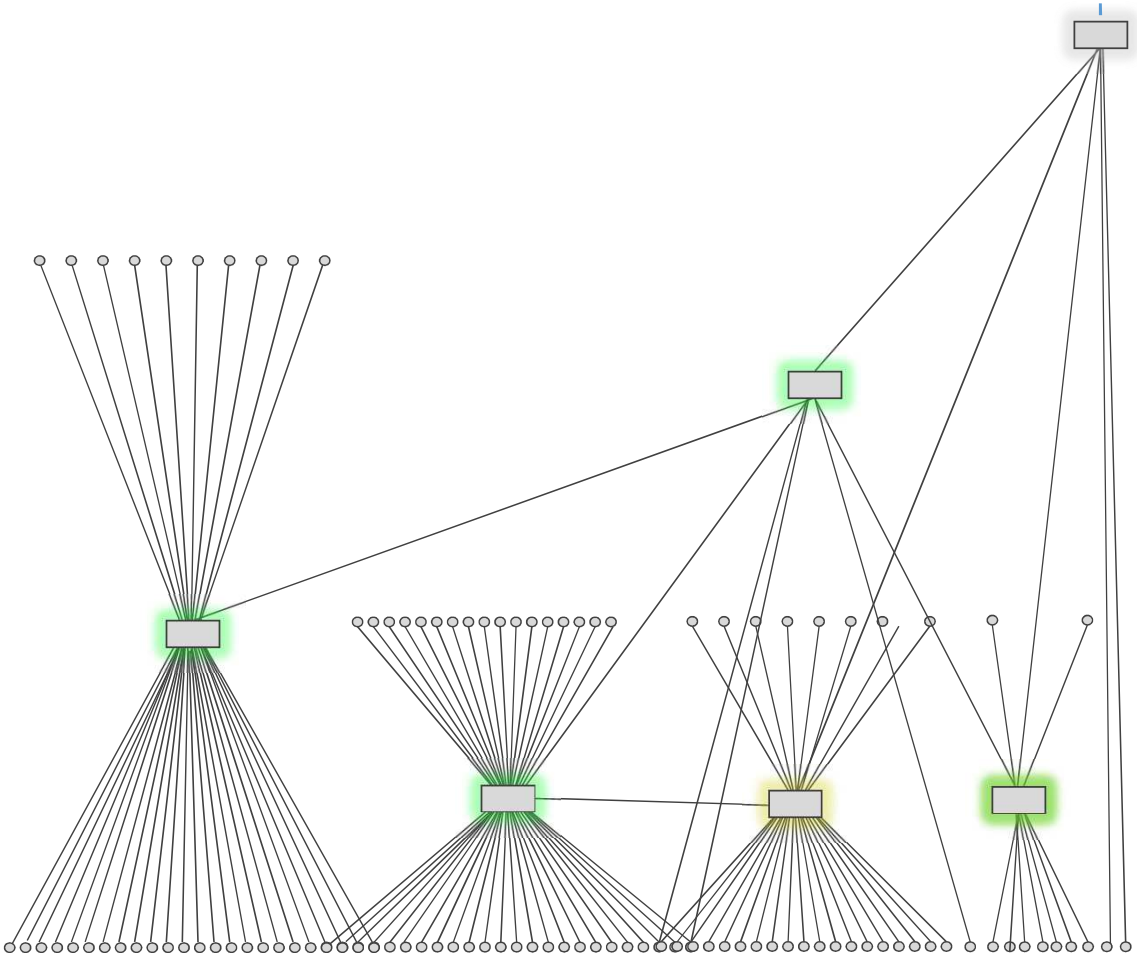
Canhos DAL, Sousa-Baena MS, de Souza S, Maia LC, Stehmann JR, et al. (2015) The Importance of Biodiversity E-infrastructures for Megadiverse Countries. PLoS Biol 13(7): e1002204. doi:10.1371/journal.pbio.1002204
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002204

Lesson learned

- The social network is an essential component of an *e*-infrastructure
- Interaction among data providers, users, and information management team is the center of innovation



… paving the way to the future



Thank you

Dora Ann Lange Canhos

dora@cria.org.br

speciesLink network - http://splink.org.br