# Latin America and the Caribbean
# Scientific Data Management Workshop

17-18 April 2018

Brazilian Academy of Sciences, Rio de Janeiro, Brazil

©Museum of Tomorrow

# PPBio’s Metacat Data Repository

- Tim Vincent
- Instituto Nacional de Pesquisas da Amazonia
- CENBAM
- PPBio

INPA
INSTITUTO NACIONAL DE
PESQUISAS DA AMAZÔNIA

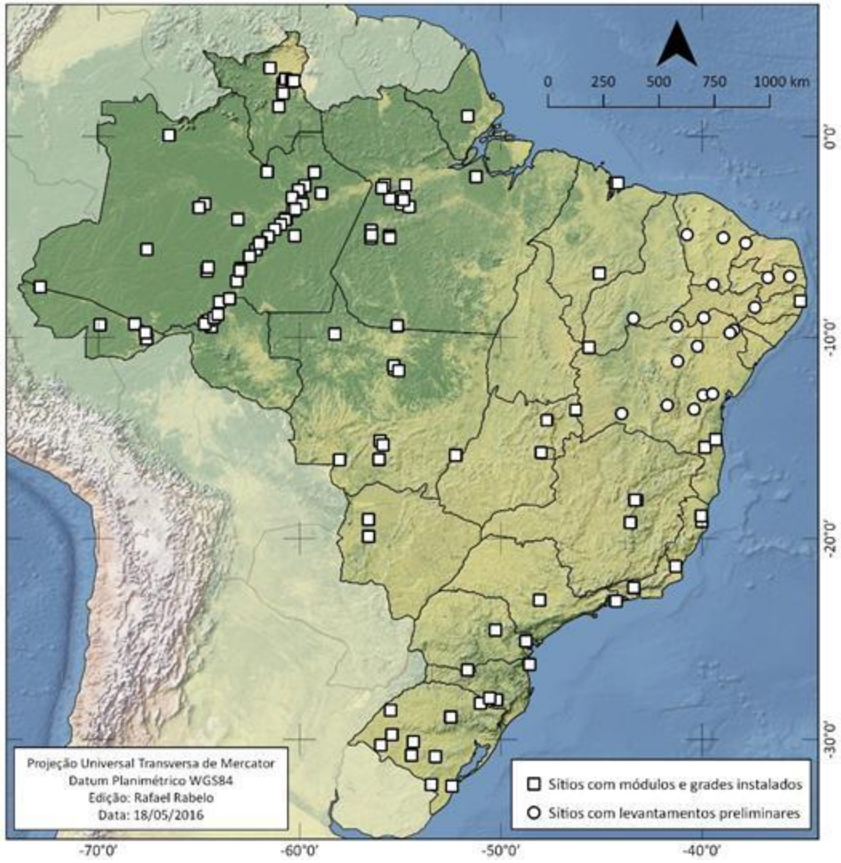
CENBAM
Centro de Estudos Integrados da
Biodiversidade Amazônica

PPBio
Programa de Pesquisa em Biodiversidade
Brazilian Biodiversity Research Programme

# PPBio

- The Biodiversity Research Program was created in 2004 with the aims of furthering biodiversity studies in Brazil, decentralizing scientific production from already-developed academic centers, integrating research activities and disseminating results across a variety of purposes, including environmental management and education.

- The Executive Nucleus of the Western Amazon Biodiversity Research Program (PPBio-AmOc) maintains a data repository used by researchers involved with several regional program cores and some other executing nuclei.



# Why did PPBio implement a Metacat biodiversity data repository?

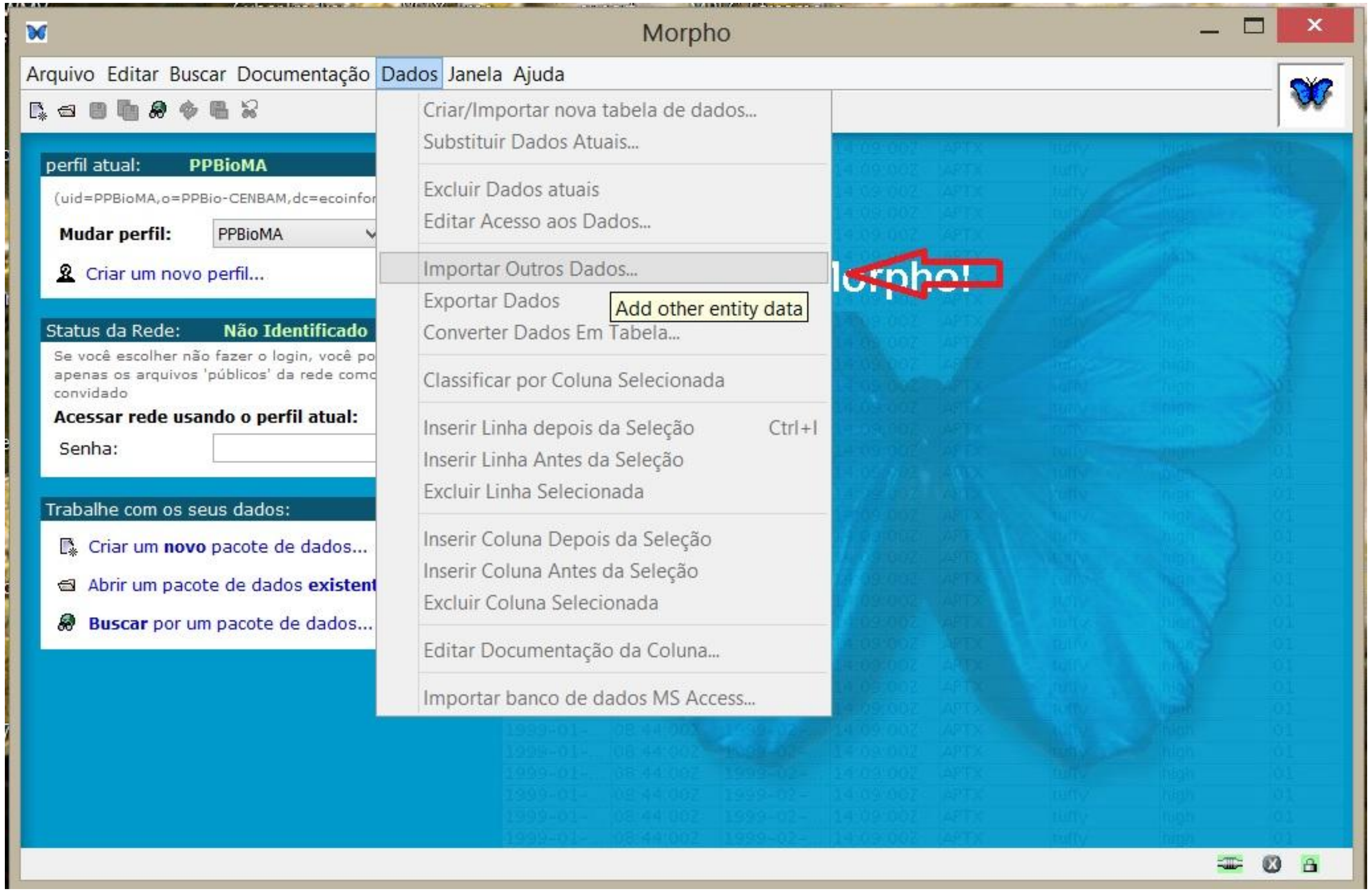
- Although this is explained in some detail in chapter 7 of the book, “Biodiversity and Integrated Environmental Monitoring” by William Magnusson et al. (2013), some of the important points are listed below:
- The most important thing to make available is the metadata.
- Cost
- Easy to upload and download. Data is in a common format.
- Ability to review and update the data package (metadata and data).

# SiBBr

- The repository’s dataset are collected and also made available on the Brazilian Biodiversity Information System (SiBBr).

- SiBBr’s Ecological data repository also uses Metacat.

Unlike worksheet style data repositories, Metacat allows a wide variety of information to be uploaded; maps, scans of raw data such as field notes, photographs and so on.



perfil atual: PPBioMA

(uid=PPBioMA,o=PPBio-CENBAM,dc=ecoinfor

Mudar perfil: PPBioMA

Status da Rede: Não Identificado

Se você escolher não fazer o login, você po
apenas os arquivos 'públicos' da rede como
convidado

Acessar rede usando o perfil atual:

Senha:

Morpho
Arquivo Editar Buscar Documentação Dados Janela Ajuda
Criar/Importar nova tabela de dados...
Substituir Dados Atuais...
Excluir Dados atuais
Editar Acesso aos Dados...
Importar Outros Dados...
Exportar Dados
Add other entity data
Converter Dados Em Tabela...
Classificar por Coluna Selecionada
Inserir Linha depois da Seleção Ctrl+I
Inserir Linha Antes da Seleção
Excluir Linha Selecionada
Inserir Coluna Depois da Seleção
Inserir Coluna Antes da Seleção
Excluir Coluna Selecionada
Editar Documentação da Coluna...
Importar banco de dados MS Access...
Criar um novo perfil...
Trabalhe com os seus dados:
Criar um novo pacote de dados...
Abrir um pacote de dados existent
Buscar por um pacote de dados...
lorpho!

# Morpho

Morpho is the Java based application that we use for creating, uploading and editing Metacat datasets.

# Morpho

Morpho is the Java based application that we use for creating, uploading and editing Metacat datasets.

Morpho is used for organizing and preparing the metadata and for uploading the dataset to the Metacat server. It can also used for accessing the dataset and updating or adding information at a later date or for making controlled access data publicly available.

This standalone tool is downloaded to the researchers own computers where the metadata and data can be prepared without the need for an internet connection.

# Morpho

Morpho is the Java based application that we use for creating, uploading and editing Metacat datasets.

Researchers use Morpho to prepare a dataset which they then send to a data manager who is responsible to checking that the metadata has been entered correctly and that there are no issues with the format of the data before uploading to the server.

This human intervention in the process is very important for quality control. Ensuring that metadata is entered correctly is important as it can influence how easily the data-set can be discovered in the future.

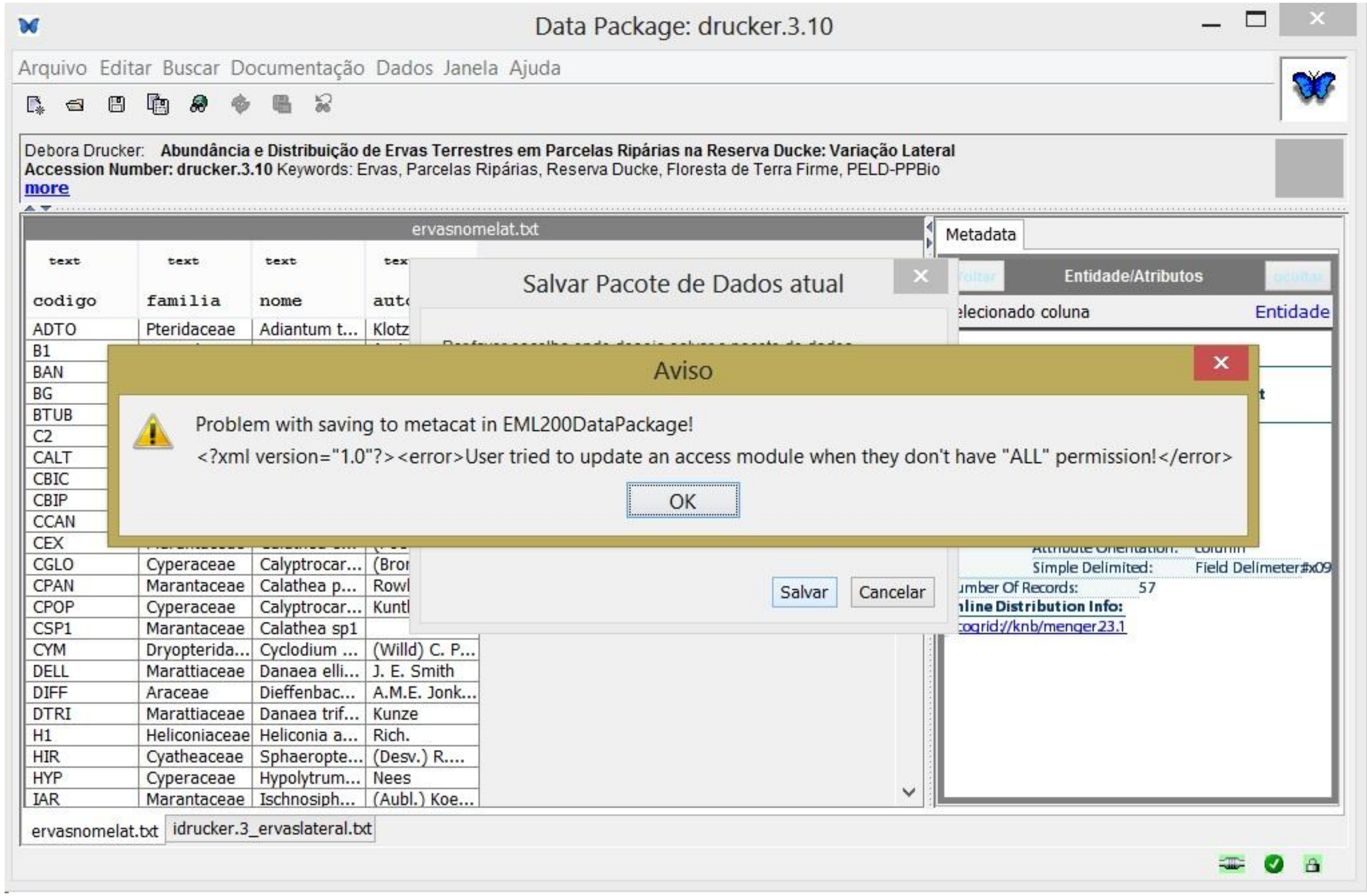
# Morpho

Morpho is the Java based application that we use for creating, uploading and editing Metacat datasets.

Natural- language metadata may describe data using an ad-hoc set of descriptive terms but there may be subsequent issues with recall. Morpho enables the data to be described using terms that facilitate retrieval, but this must be correctly done by the person entering the data.

The PPBio website provides detailed instructions on how to use Morpho, but person-to- person training is very useful since there are some small hacks that are necessary in order to resolve issues with Morpho's user interface and the terminology used for defining and describing the data.

# When there are problems however, the help you get from the NCEAS/KNB team is very good.


Data Package: drucker.3.10
Arquivo Editar Buscar Documentação Dados Janela Ajuda
more
ervasnomelat.txt
Metadata
Voltar
Entidade/Atributos
ocultar
Salvar Pacote de Dados atual
Aviso
OK
Salvar
Cancelar
ervasnomelat.txt
idrucker.3_ervaslateral.txt

Debora Drucker: **Abundância e Distribuição de Ervas Terrestres em Parcelas Ripárias na Reserva Ducke: Variação Lateral**
**Accession Number: drucker.3.10** Keywords: Ervas, Parcelas Ripárias, Reserva Ducke, Floresta de Terra Firme, PELD-PPBio

| text | text | text | tex |
|---|---|---|---|
| codigo | familia | nome | aut |
| ADTO | Pteridaceae | Adiantum t... | Klotz |
| B1 | | | |
| BAN | | | |
| BG | | | |
| BTUB | | | |
| C2 | | | |
| CALT | | | |
| CBIC | | | |
| CBIP | | | |
| CCAN | | | |
| CEX | | | |
| CGLO | Cyperaceae | Calyptrocar... | (Bror |
| CPAN | Marantaceae | Calathea p... | Rowl |
| CPOP | Cyperaceae | Calyptrocar... | Kuntl |
| CSP1 | Marantaceae | Calathea sp1 | |
| CYM | Dryopterida... | Cyclodium ... | (Willd) C. P... |
| DELL | Marattiaceae | Danaea elli... | J. E. Smith |
| DIFF | Araceae | Dieffenbac... | A.M.E. Jonk... |
| DTRI | Marattiaceae | Danaea trif... | Kunze |
| H1 | Heliconiaceae | Heliconia a... | Rich. |
| HIR | Cyatheaceae | Sphaeropte... | (Desv.) R.... |
| HYP | Cyperaceae | Hypolytrum... | Nees |
| IAR | Marantaceae | Ischnosiph... | (Aubl.) Koe... |

```
Problem with saving to metacat in EML200DataPackage!
<?xml version="1.0"?><error>User tried to update an access module when they don't have "ALL" permission!</error>
```

elecionado coluna Entidade

Simple Delimited: Field Delimeter#x09
umber Of Records: 57
**nline Distribution Info:**
cogrid://knb/menger23.1

# NCEAS

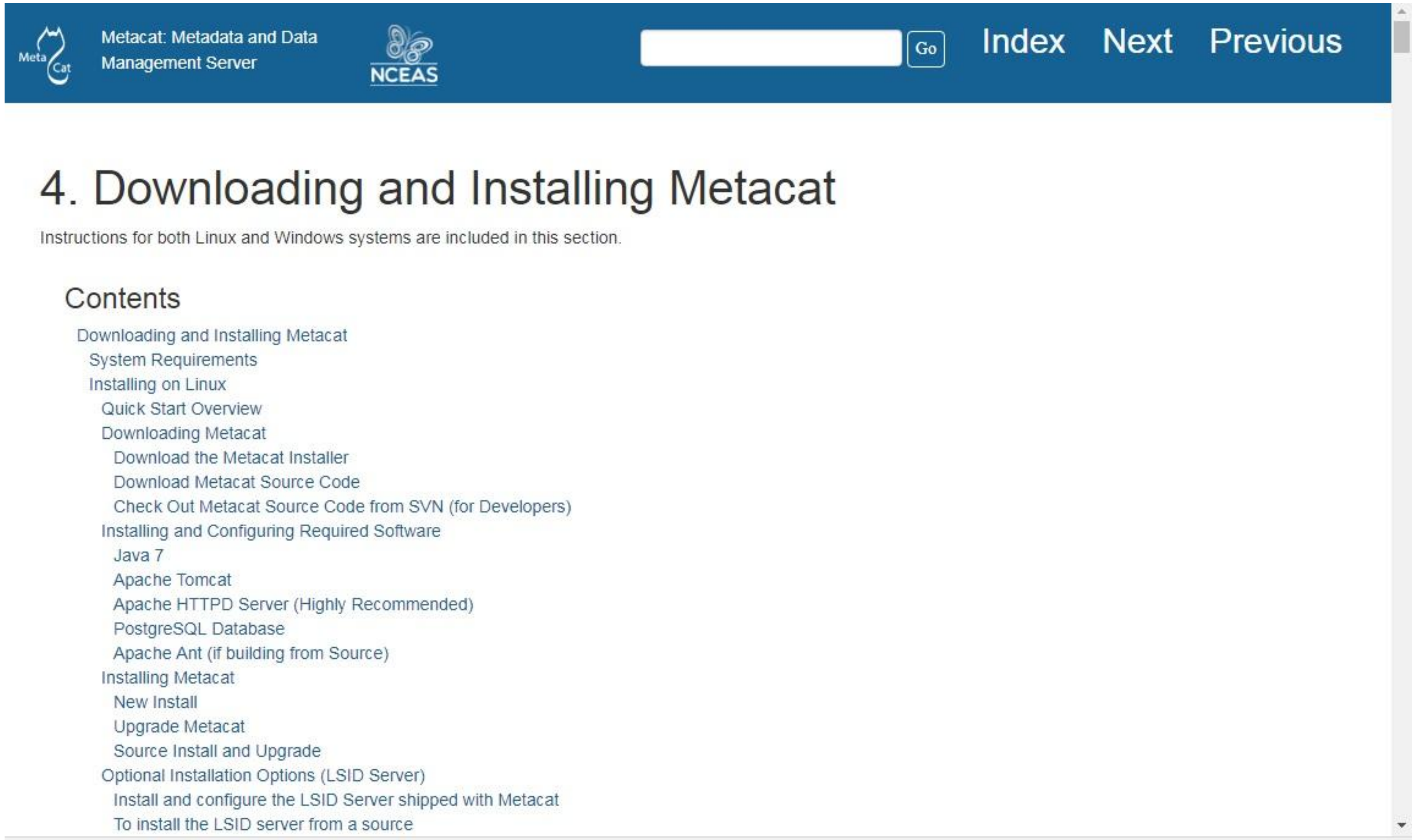
- National Center for Ecological Analysis and Synthesis
- Metacat and Morpho are products of NCEAS.

NCEAS - About

# Metacat

- “...flexible, open source metadata catalogue and data repository...”
- “...utilizes a relational database management system to store XML and associated meta-level information.”
- Technical Expertise Required: Basic programming skills.
- Cost: Free

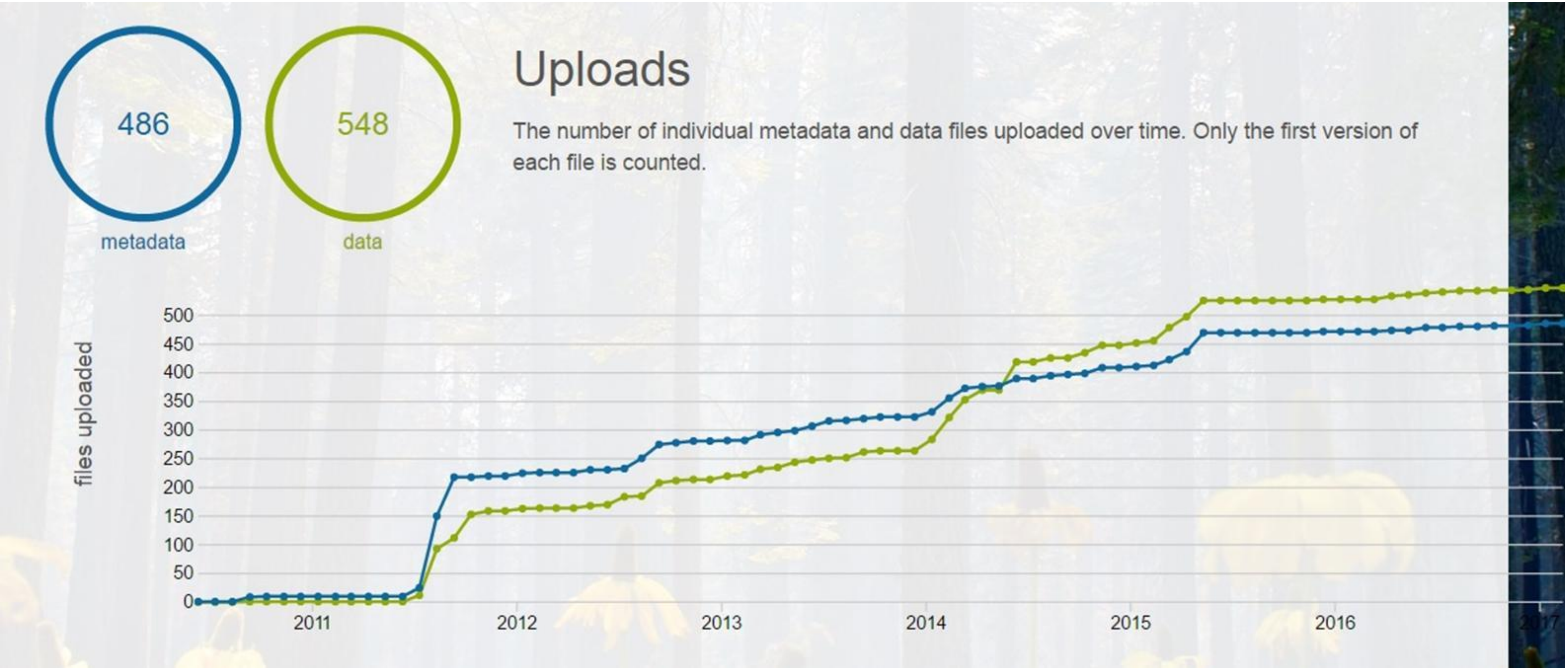
# Follow the instructions for installing Metacat...



## 4. Downloading and Installing Metacat

Instructions for both Linux and Windows systems are included in this section.

### Contents



486

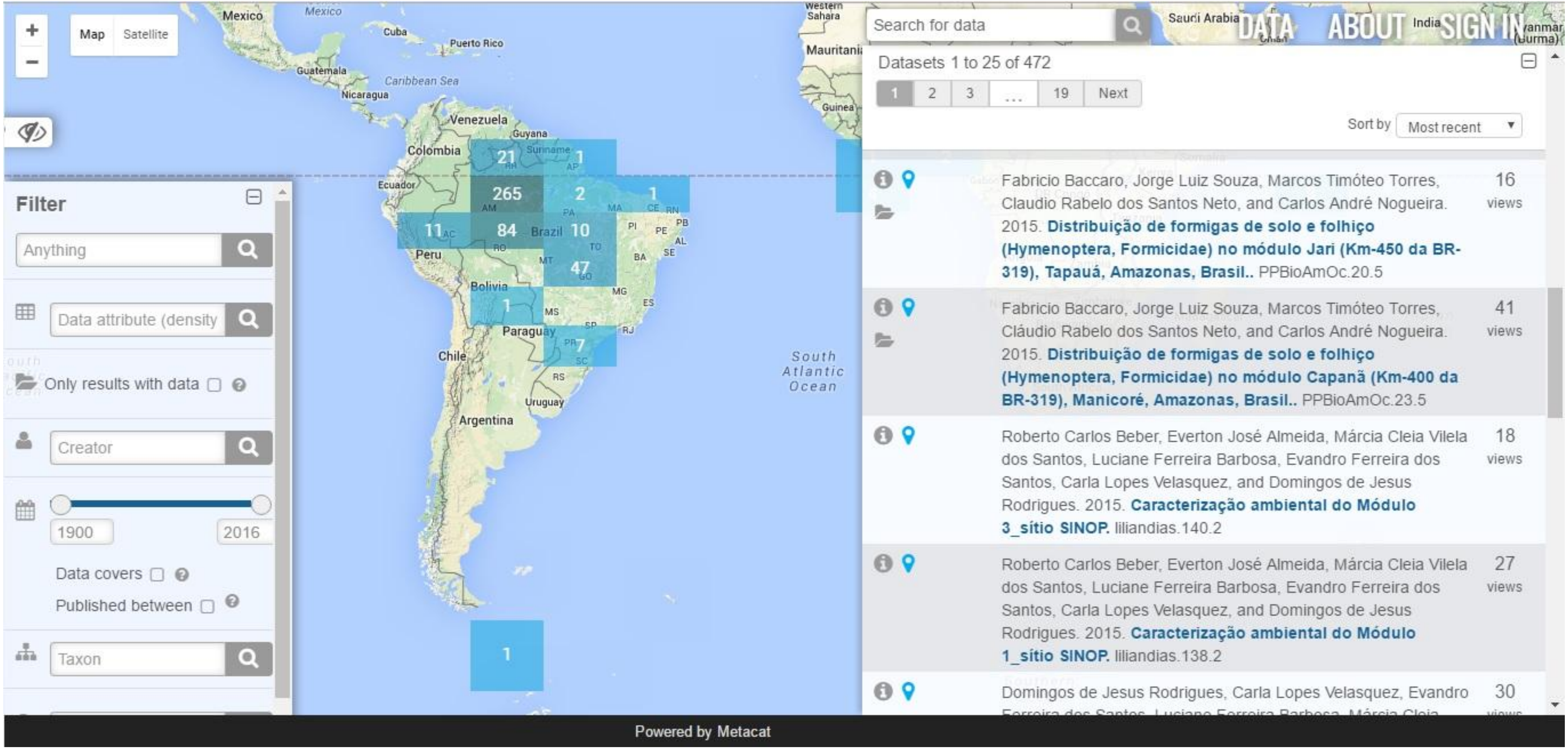
metadata

548

data

# Uploads

The number of individual metadata and data files uploaded over time. Only the first version of each file is counted.



Datasets 1 to 25 of 472

Only results with data □

Data covers □

Published between □

Fabricio Baccaro, Jorge Luiz Souza, Marcos Timóteo Torres, Claudio Rabelo dos Santos Neto, and Carlos André Nogueira. 2015. **Distribuição de formigas de solo e folhiço (Hymenoptera, Formicidae) no módulo Jari (Km-450 da BR-319), Tapauá, Amazonas, Brasil..** PPBioAmOc.20.5 — 16 views

Fabricio Baccaro, Jorge Luiz Souza, Marcos Timóteo Torres, Cláudio Rabelo dos Santos Neto, and Carlos André Nogueira. 2015. **Distribuição de formigas de solo e folhiço (Hymenoptera, Formicidae) no módulo Capanã (Km-400 da BR-319), Manicoré, Amazonas, Brasil..** PPBioAmOc.23.5 — 41 views

Roberto Carlos Beber, Everton José Almeida, Márcia Cleia Vilela dos Santos, Luciane Ferreira Barbosa, Evandro Ferreira dos Santos, Carla Lopes Velasquez, and Domingos de Jesus Rodrigues. 2015. **Caracterização ambiental do Módulo 3_sítio SINOP.** liliandias.140.2 — 18 views

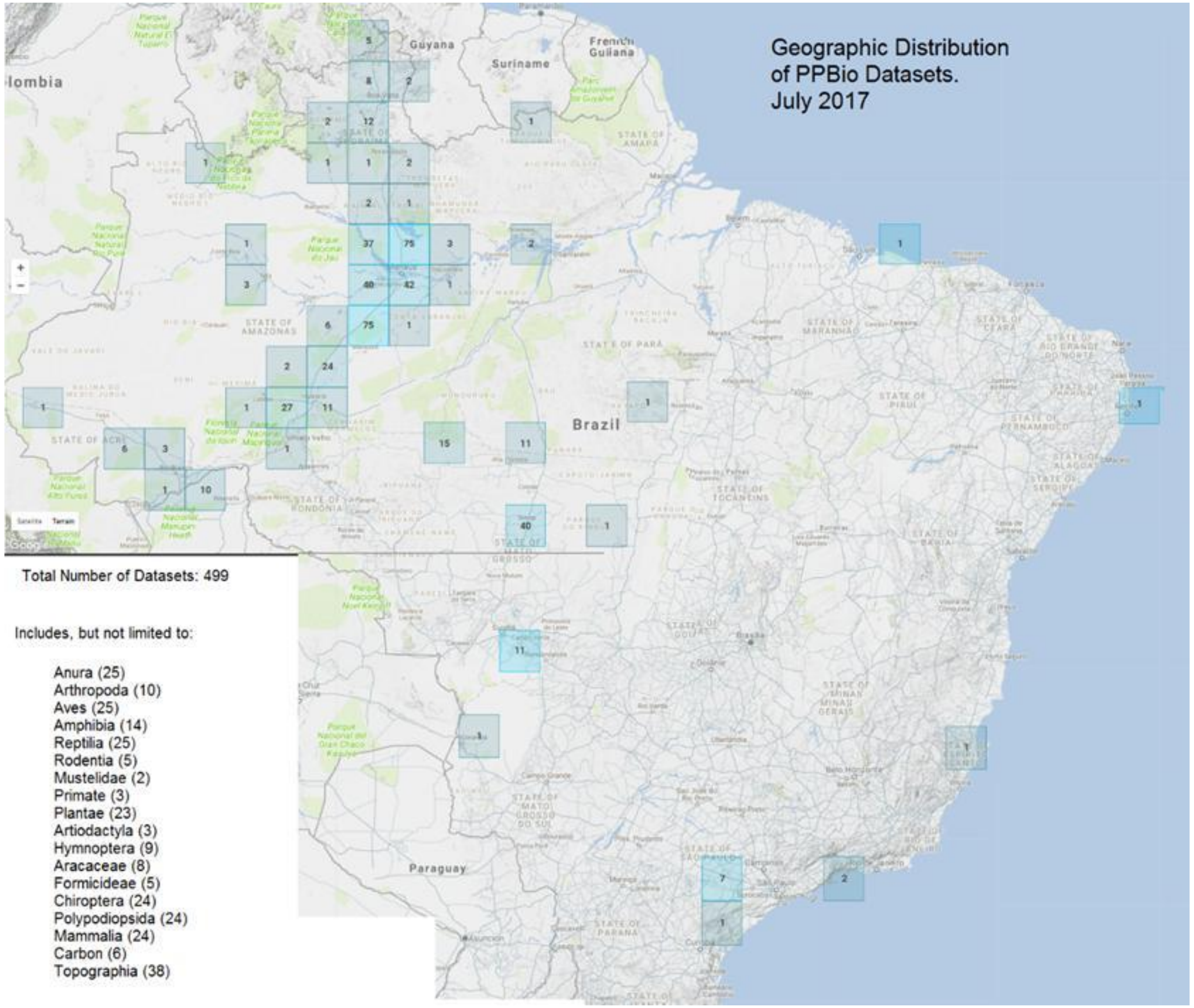
Roberto Carlos Beber, Everton José Almeida, Márcia Cleia Vilela dos Santos, Luciane Ferreira Barbosa, Evandro Ferreira dos Santos, Carla Lopes Velasquez, and Domingos de Jesus Rodrigues. 2015. **Caracterização ambiental do Módulo 1_sítio SINOP.** liliandias.138.2 — 27 views

Domingos de Jesus Rodrigues, Carla Lopes Velasquez, Evandro — 30

# Example:

Claudia Keller, Francisco Villamarín, Rafael Bernhard, and Daniely Félix-Silva. 2016. Chelonian records from the upper Madeira River and the Madeira-Purus interfluvium, 2011- 2015.

urn:node:PPBIO. PPBioAmOc.50.30.

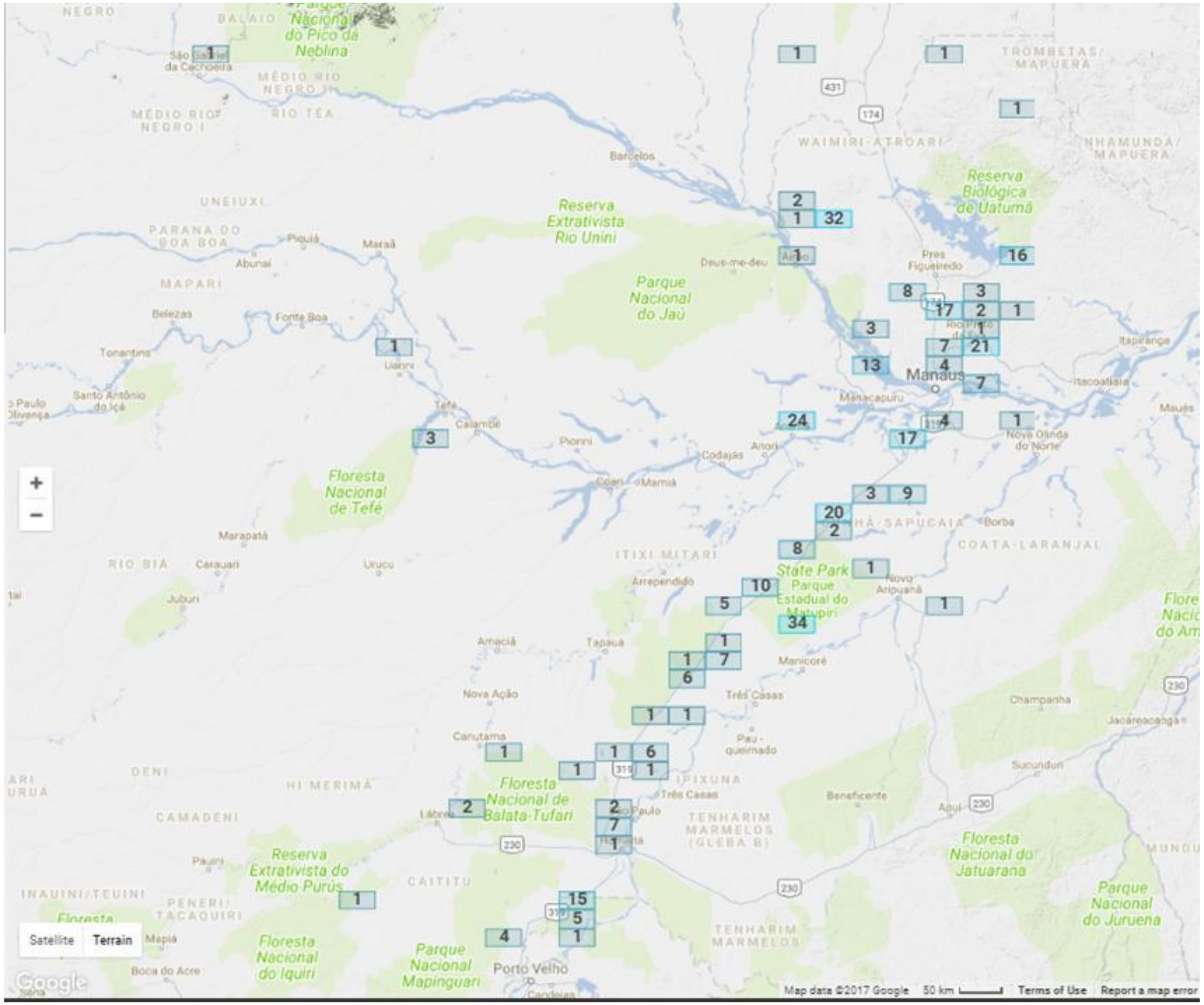


## Geographic Distribution of PPBio Datasets. July 2017

Total Number of Datasets: 499

Includes, but not limited to:

- Anura (25)
- Arthropoda (10)
- Aves (25)
- Amphibia (14)
- Reptilia (25)
- Rodentia (5)
- Mustelidae (2)
- Primate (3)
- Plantae (23)
- Artiodactyla (3)
- Hymnoptera (9)
- Aracaceae (8)
- Formicideae (5)
- Chiroptera (24)
- Polypodiopsida (24)
- Mammalia (24)
- Carbon (6)
- Topographia (38)


NEGRO
BALAIO
Parque Nacional do Pico da Neblina
São Gabriel da Cachoeira
1
MÉDIO RIO NEGRO II
MÉDIO RIO NEGRO I
RIO TÉA
1
431
174
1
TROMBETAS/MAPUERA
1
WAIMIRI-ATROARI
NHAMUNDÁ/MAPUERA
Barcelos
Reserva Extrativista Rio Unini
Reserva Biológica de Uatumã
UNEIUXI
2
1
32
PARANA DO BOA BOA
Piquiá
Maraã
Abunaí
Deus-me-deu
Airão
1
Pres. Figueiredo
16
MAPARI
Parque Nacional do Jaú
8
3
17
2
1
Belezas
Fonte Boa
3
Rio Preto
1
7
21
Tonantins
1
Uarini
13
4
Manaus
7
Itapiranga
Itacoatiara
São Paulo de Olivença
Santo Antônio do Içá
Tefé
Manacapuru
Maués
3
Caiambé
Pioroni
Codajás
Anori
24
4
1
17
Nova Olinda do Norte
Coari
Mamiá
Floresta Nacional de Tefé
3
9
20
2
Borba
Marapatá
MATÁ-SAPUCAIA
COATA-LARANJAL
8
ITIXI MITARI
State Park
1
RIO BIÁ
Carauari
Urucu
Arrependido
Parque Estadual do Matupiri
10
Novo Aripuanã
Juburi
5
34
1
Floresta Nacional do Amazonas
Amaciá
Tapauá
1
7
6
Manicoré
Nova Ação
Três Casas
230
Champanha
1
1
Jacareacanga
Canutama
Pau-queimado
1
1
6
1
1
319
Sucunduri
DENI
HI MERIMÃ
IPIXUNA
Floresta Nacional de Balata-Tufari
2
2
Humaitá
7
1
Três Casas
Beneficente
Apuí
230
CAMADENI
TENHARIM MARMELOS (GLEBA B)
Lábrea
230
MUNDU
Floresta Nacional do Jatuarana
Pauini
Reserva Extrativista do Médio Purus
CAITITU
230
Parque Nacional do Juruena
INAUINI/TEUINI
PENERI/TACAQUIRI
1
15
Floresta
5
319
TENHARIM MARMELOS
Mapiá
4
1
Floresta Nacional do Iquiri
Parque Nacional Mapinguari
Porto Velho
Boca do Acre
Candeias
+
−
Satellite
Terrain
Google
Map data ©2017 Google
50 km
Terms of Use
Report a map error

PPBio-AmOc is a node of the Earth Data Observation Network (DataONE).

DataONE

Data Observation Network for Earth

Data Observation Network for Earth (DataONE) is the foundation of new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data.

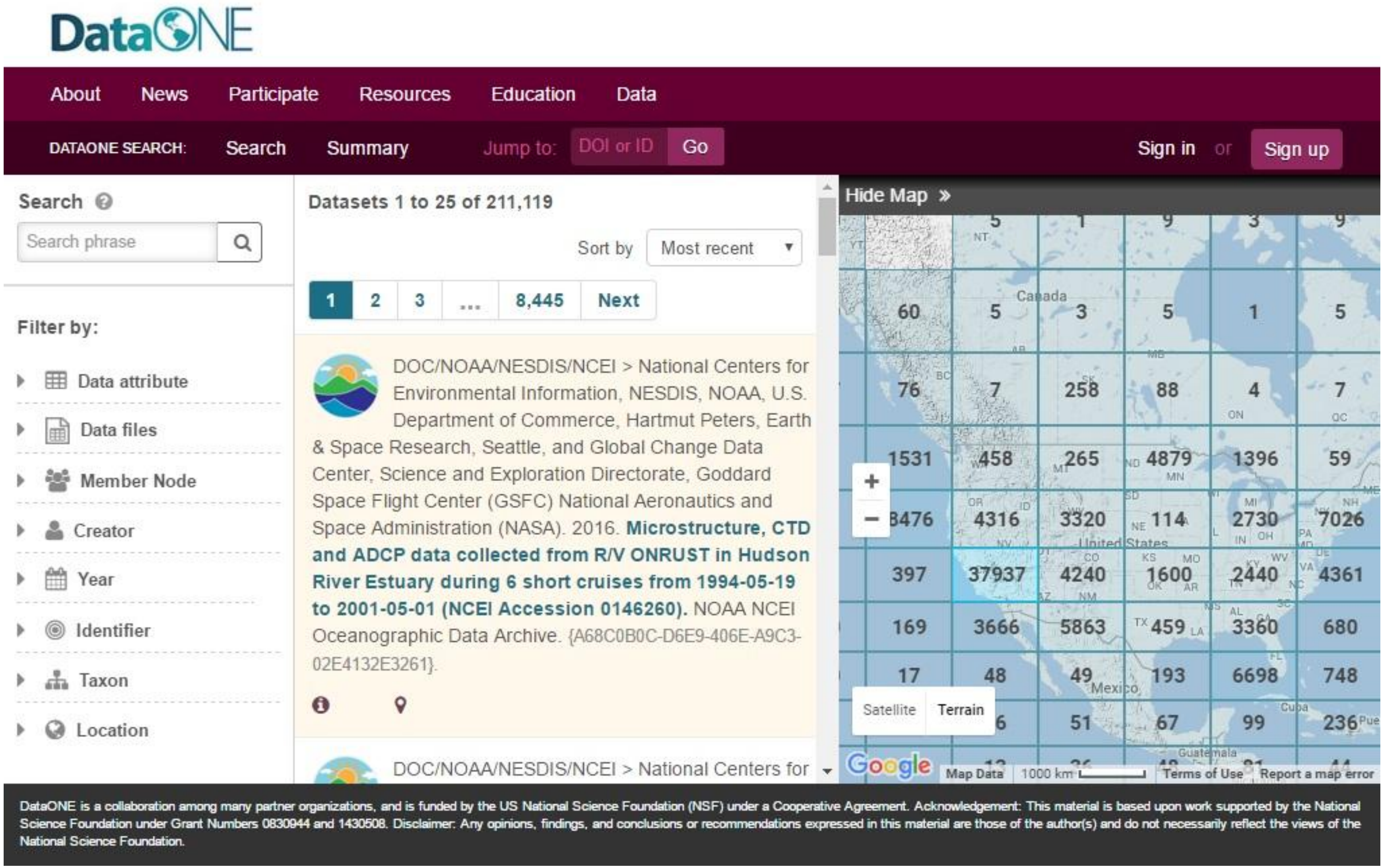
DataONE
Data Observation Network for Earth

DataONE will ensure the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national science data via three primary cyberinfrastucture elements and a broad education and outreach program.

DataONE comprises a distributed network of data centers, science networks or organizations. These organizations can expose their data within the DataONE network through the implementation of the DataONE Member Node service interface. In addition to scientific data, Member Nodes can provide computing resources, or services such as data replication, to the DataONE community.

Member Nodes

# DataOne GUI



Datasets 1 to 25 of 211,119

DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce, Hartmut Peters, Earth & Space Research, Seattle, and Global Change Data Center, Science and Exploration Directorate, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA). 2016. **Microstructure, CTD and ADCP data collected from R/V ONRUST in Hudson River Estuary during 6 short cruises from 1994-05-19 to 2001-05-01 (NCEI Accession 0146260).** NOAA NCEI Oceanographic Data Archive. {A68C0B0C-D6E9-406E-A9C3-02E4132E3261}.

DOC/NOAA/NESDIS/NCEI > National Centers for

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

# Deliverable 2.2 (D2.2) Data sharing tools

- Metacat, Morpho and DataOne were included in the February 2106 EUBON publication “Data sharing tools”.
- The paper gives a comprehensive review of the existing tools for metadata, occurrence data, and ecological data.
- It includes a detailed description of the tools, their pros and cons, is followed by recommendations on their deployment and enhancement.
- Available online at
- http://www.eubon.eu/news/13351_D2.2 Data sharing tools

# Search Metacat using Solr

- Metacat uses the Solr open source search platform to organise and index the metadata and data contained in the database and to support search operations.
- This means that anyone can query a public data repository such as ours and get information according to their specific requirements.
- A Solr query begins with a base URL , in our case, this is "https://ppbiodata.inpa.gov.br/metacat/d1/mn/v2/query/solr/" which is followed by a specific set of questions, followed by the output that is required.

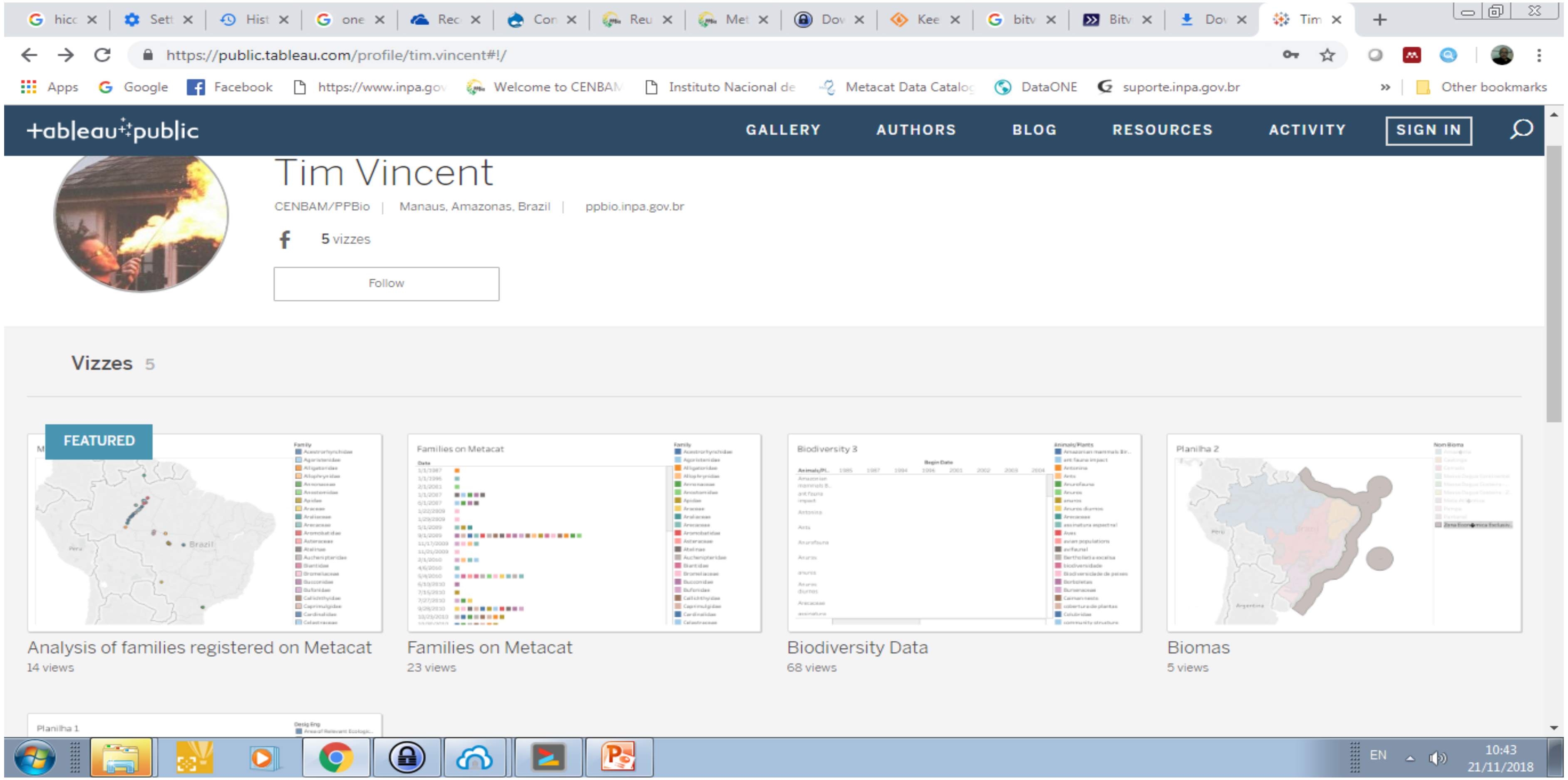
# Search Metacat using Solr

- So,the questions for example might ask for records that contain any kind of taxonomic information, for example ?q=class:* OR family:* OR genus:* OR kingdom:* OR order:* OR phylum:* OR scientificName:* OR species:*
- and the output takes the following format, which includes the ID number of the record, the geographic boundaries, taxomonic information related tto the record and the time that the information was collected:
- &fl=identifier,northBoundCoord,eastBoundCoord,southBoundCoord,westBoundCoord,class,family,genus,kingdom,order,phylum,scientificName,species,beginDate,
- &rows=1000 (the default value is 10) - You might need to increase this if your output file has exactly 1000 rows!
- &wt=csv indicates a request for the output to be in csv format.

# Search Metacat using Solr

- So, the complete URL for the query is:
- https:// ppbiodata.inpa.gov.br/metacat/d1/mn/v2/query/solr/?q=class:* OR family:* OR genus:* OR kingdom:* OR order:* OR phylum:* OR scientificName:* OR species:* &fl=identifier,northBoundCoord,eastBoundCoord,southBoundCoord, westBoundCoord,class,family,genus,kingdom,order,phylum,scientificName,species,beginDate,dateModified &rows=1000 &wt=csv
- Metacat and Tableau - Using the free application Tableau, data can be visualised in a variety of different ways.

# Tableau



https://public.tableau.com/profile/tim.vincent#!/

Tim Vincent

CENBAM/PPBio | Manaus, Amazonas, Brazil | ppbio.inpa.gov.br

5 vizzes

Vizzes 5

Analysis of families registered on Metacat
14 views

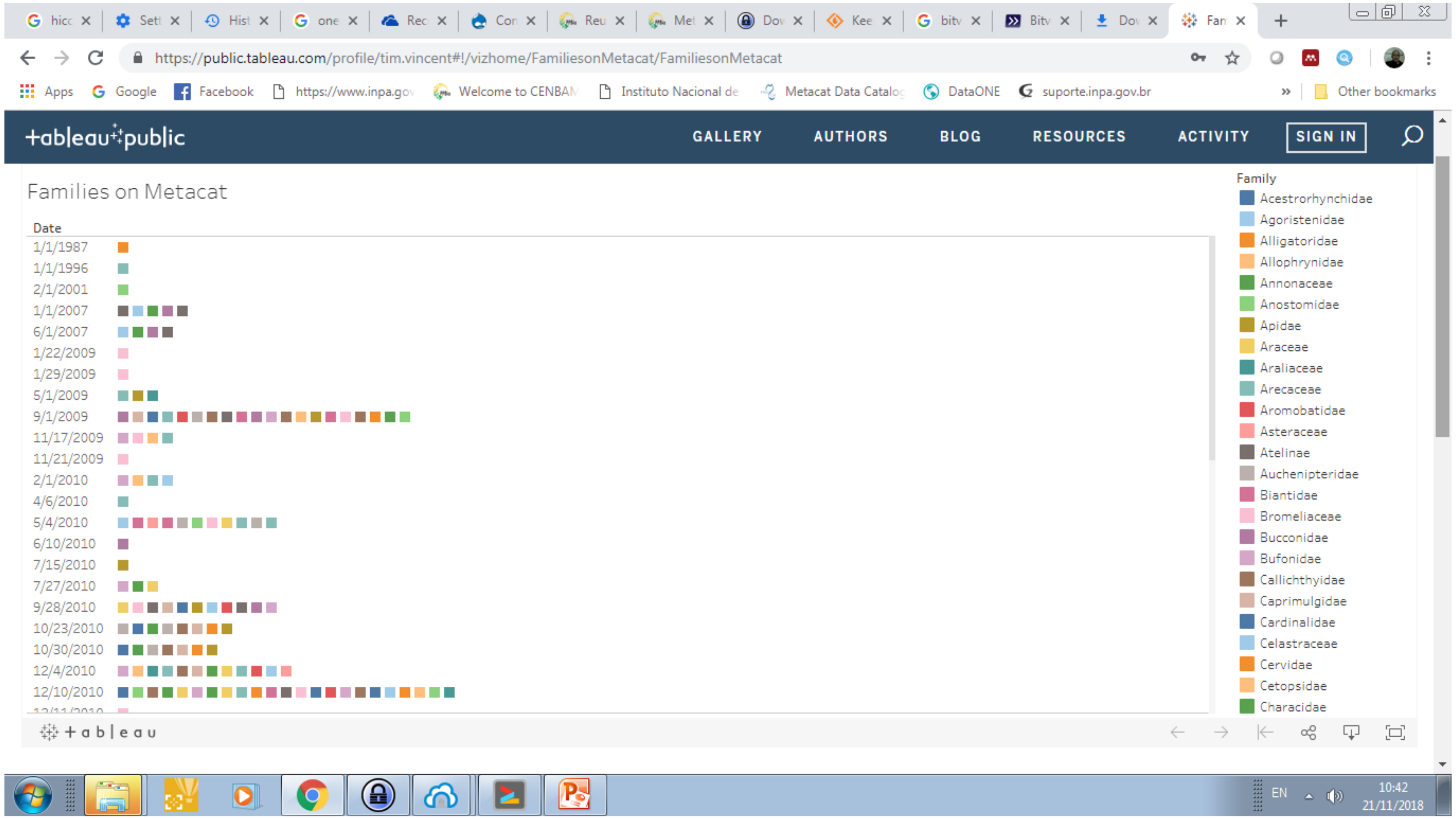
Families on Metacat
23 views

Biodiversity Data
68 views

Biomas
5 views

Planilha 1

# Tableau



Families on Metacat

Date

1/1/1987
1/1/1996
2/1/2001
1/1/2007
6/1/2007
1/22/2009
1/29/2009
5/1/2009
9/1/2009
11/17/2009
11/21/2009
2/1/2010
4/6/2010
5/4/2010
6/10/2010
7/15/2010
7/27/2010
9/28/2010
10/23/2010
10/30/2010
12/4/2010
12/10/2010

Family

Acestrorhynchidae
Agoristenidae
Alligatoridae
Allophrynidae
Annonaceae
Anostomidae
Apidae
Araceae
Araliaceae
Arecaceae
Aromobatidae
Asteraceae
Atelinae
Auchenipteridae
Biantidae
Bromeliaceae
Bucconidae
Bufonidae
Callichthyidae
Caprimulgidae
Cardinalidae
Celastraceae
Cervidae
Cetopsidae
Characidae

# Certification of repositories