Estratégias para armazenamento de dados biológicos e ambientais: experiências do PELD/LTER e PPBio

Flávia Pezzini (INPA/PPBio/PELD Banco de Dados)
Flávia Costa (INPA/PPBio/PELD Amazônia)
William Magnusson (INPA/PPBio)

Bancos x Repositórios

- Bancos de dados são específicos para cada tipo de dado

- Pode ser relativamente simples montar e manejar um banco de dados quando o tipo de dado é simples e não muda com o tempo
  – p. ex. GENBANK ou bancos de dados de coleções de espécimes

Armazenamento de dados de coletas biológicas

- Já existem muitas inciativas para armazenar dados de coletas biológicas
- Várias coleções (Herbários, Museus) tem feito esforços para colocar suas informações em bancos padronizados e que foram adotados por consenso (p.ex BRAHMS e Specify)

Dados complexos

- Dados de inventários biológicos são mais variáveis e precisam de abordagens mais específicas.

- É muito difícil prever todos os possíveis tipos de dados e estrutura das amostragens a priori, e portanto, definir uma estrutura de banco de dados que funcione para tudo

How to Manage Data Badly (Part 1)

In a landmark article in *The American Statistician*, Howard Wainer (1984) presented ideas for "How to Display Data Badly," wherein good data are ruined by bad graphics. Wainer presumed too much. In this essay, I extend his concept by presenting ideas and examples of how scientific data can be managed badly so that they never even make it to the graphics stage. Modern database management software, continually improving hardware and networks, and many sound recommendations for managing ecological data (e.g., ESA 1995, NRC 1995) are making it increasingly difficult to manage data badly. It can still be done, however, and judging by various so-called "horror stories" one hears about adventures with ecological and other scientific databases, is done frequently. Those people still having trouble mismanaging data, whether they are database managers, administrators, or scientists, will find the following techniques helpful.

Techniques for database managers

Rule 1. "One world, one

Two time-honored te
by database managers ar

hard to get data into the system, and (2) make it hard to get data out of the system. Although experts often use both techniques, novices may wish to start with just one. A data system that is far more complex than is necessary will usually do the trick. In *Principles of Data-Base Management*, James Martin (1976) suggested that one reason for failure of long-term databases is "Plans for the installation of a grandiose all-embracing system." This strategy is effective because it makes development time long, data loading slow, and data queries difficult to formulate. Discouraged data

How to Manage Data Badly (Part 2)

Preamble. Although managing data badly is difficult, given new developments in hardware and software and many sound recommendations for how to manage ecological data (ESA 1995, NRC 1995), it is still being done. In the spirit of Howard Wainer's (1984) article in *The American Statistician* on "How to Display Data Badly," Part 1 of this essay gave 10 rules for those database managers and administrators who are determined to manage their data badly. Part 2 gives some tips for scientists with the same objective.

Techniques for scientists

Rule 11. Hoard your data

The best technique for scientists

data hoarders can carry their data to the grave. According to a report by the Ecological Society of America (ESA 1995), there are ". . . numerous incidences of researchers ending their careers without the resources to make provision for the curation and maintenance of their long-term data sets." Data sharing in many collaborative efforts is inhibited by ". . . unclear responsibilities, conflicting goals, misunderstandings, and outright rivalries" (NRC 1995).

Rule 12. It is better to get than to manage

Do not be misled by those who have tried to encourage slovenliness in scientists by proposing that they be rewarded for publishing high-quality data sets in the same way they are rewarded for publishing scientific papers. This could lead to the specter of a publishing person perishing. Do not

to collect they are
saving? Simple: make
saving. These next
paths to follow, both
you to avoid tedious
verify the accuracy o
(2) avoid writing met
tion about data); inst
sumptions, processi
quality of the data re
to future users.

Amusing things
data are not quality
ample, who would sus
researcher could ma
error of swapping lati
tude? Let those GIS
try to lie about why th
tion from the east co
on the Greenland ice
more effective in dr
ers than an error-ri
Celko (1994) reminds
nities to corrupt good

ESA 1995, NRC 1995) are making it increasingly difficult to manage data badly. It can still be done, however, and judging by various so-called "horror stories" one hears about adventures with ecological and other scientific databases, is done frequently. Those people still having trouble mismanaging data, whether they are database managers, administrators, or scientists, will find the following techniques helpful.

Techniques for database managers

Rule 1. "One world, one database"

Two time-honored techniques used by database managers are: (1) make it hard to get data into the system, and (2) make it hard to get data out of the system. Although experts often use both techniques, novices may wish to start with just one. A data system that is far more complex than is necessary will usually do the trick. In *Principles of Data-Base Management*, James Martin (1976) suggested that one reason for failure of long-term databases is "Plans for the installation of a grandiose all-embracing system." This strategy is effective because it makes development time long, data loading slow, and data queries difficult to formulate. Discouraged data collectors and users will seek solutions elsewhere. In one case, a national marine water quality database

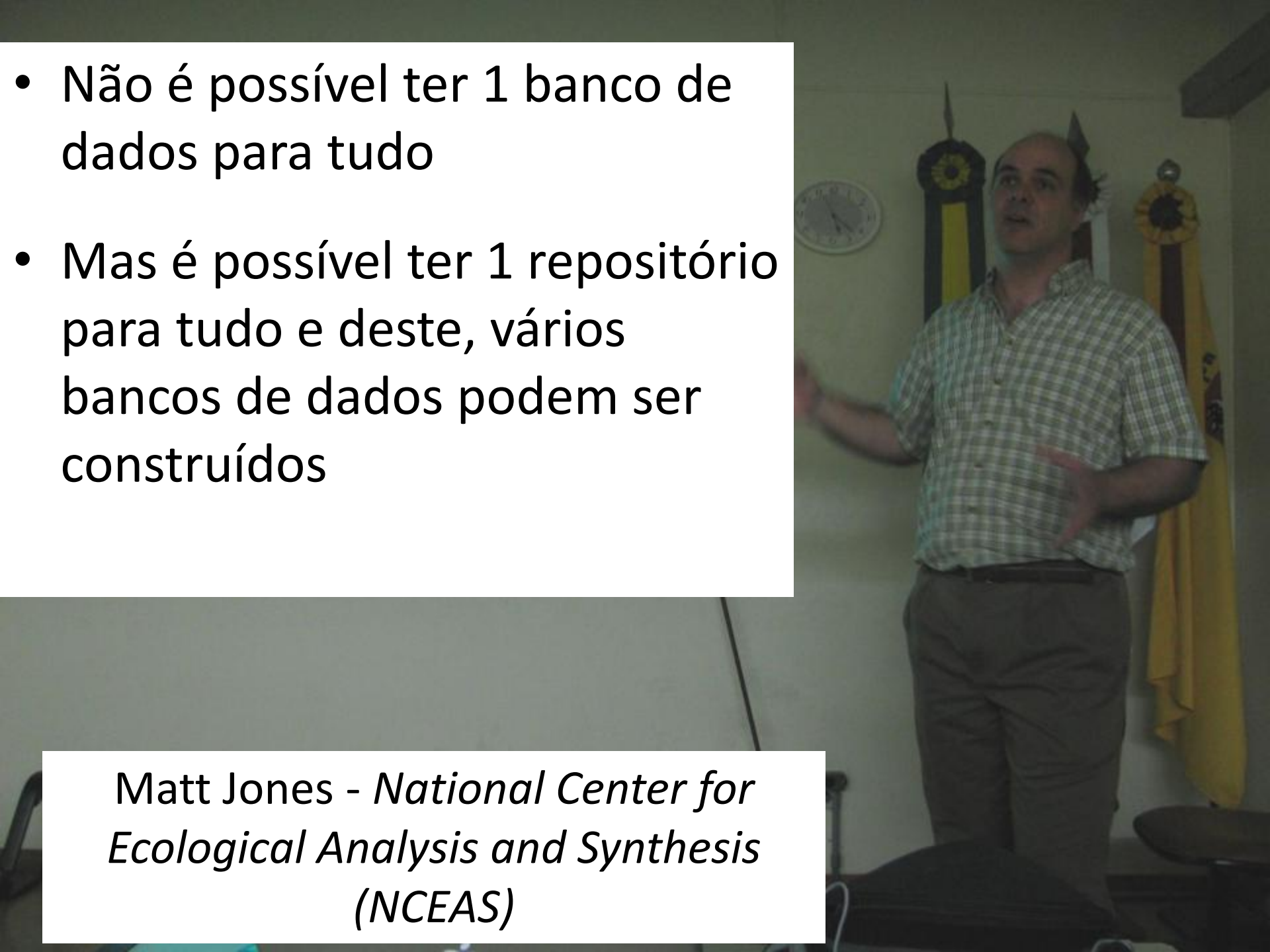
NATURE|Vol 461|10 September 2009

Empty archives

Most researchers agree that open access to data is the scientific ideal, so what is stopping it happening? **Bryn Nelson** investigates why many researchers choose not to share.

Figura: Nature: 461 (7261), 2010.

- Não é possível ter 1 banco de dados para tudo
- Mas é possível ter 1 repositório para tudo e deste, vários bancos de dados podem ser construídos



Matt Jones - *National Center for Ecological Analysis and Synthesis (NCEAS)*

Repositórios

Metadados são informações que descrevem os dados

e sem eles, os dados não tem nenhum valor ou utilidade, seja lá qual for o sistema de armazenagem.

Data Manager Library

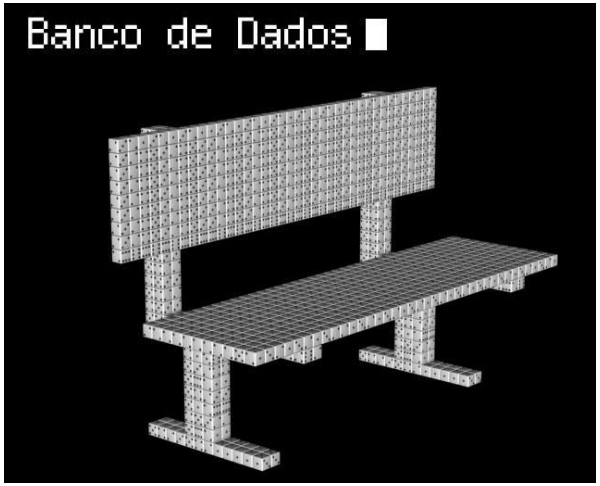
"Provide metadata-based query access to data with the ability to filter, join, and concatenate across heterogeneous data sets."

Biblioteca de Manejo de Dados

"Proporcionar consultas aos dados baseada nos metadados, com a capacidade de filtrar, juntar e concatenar através de conjuntos de dados heterogêneos."

METADADOS

BIBLIOTECA DE DADOS



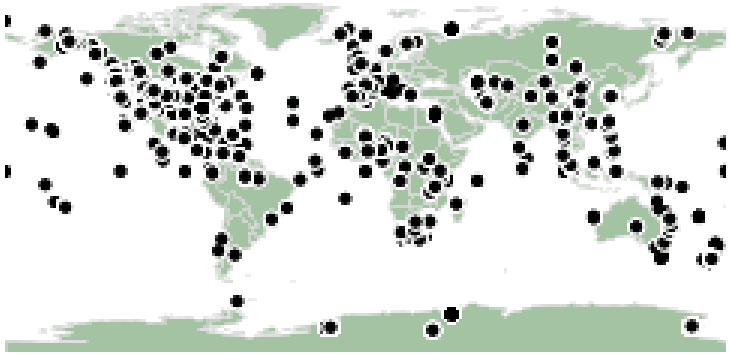
BANCO DE DADOS

Linguagem para Metadados

- Já foram desenvolvidas linguagens e modelos para metadados de várias disciplinas
- para dados de inventários biológicos e ambientais, um bom modelo é o

Ecological Metadada Language (EML) desenvolvido pelo Knowledge Network for Biocomplexity (KNB)

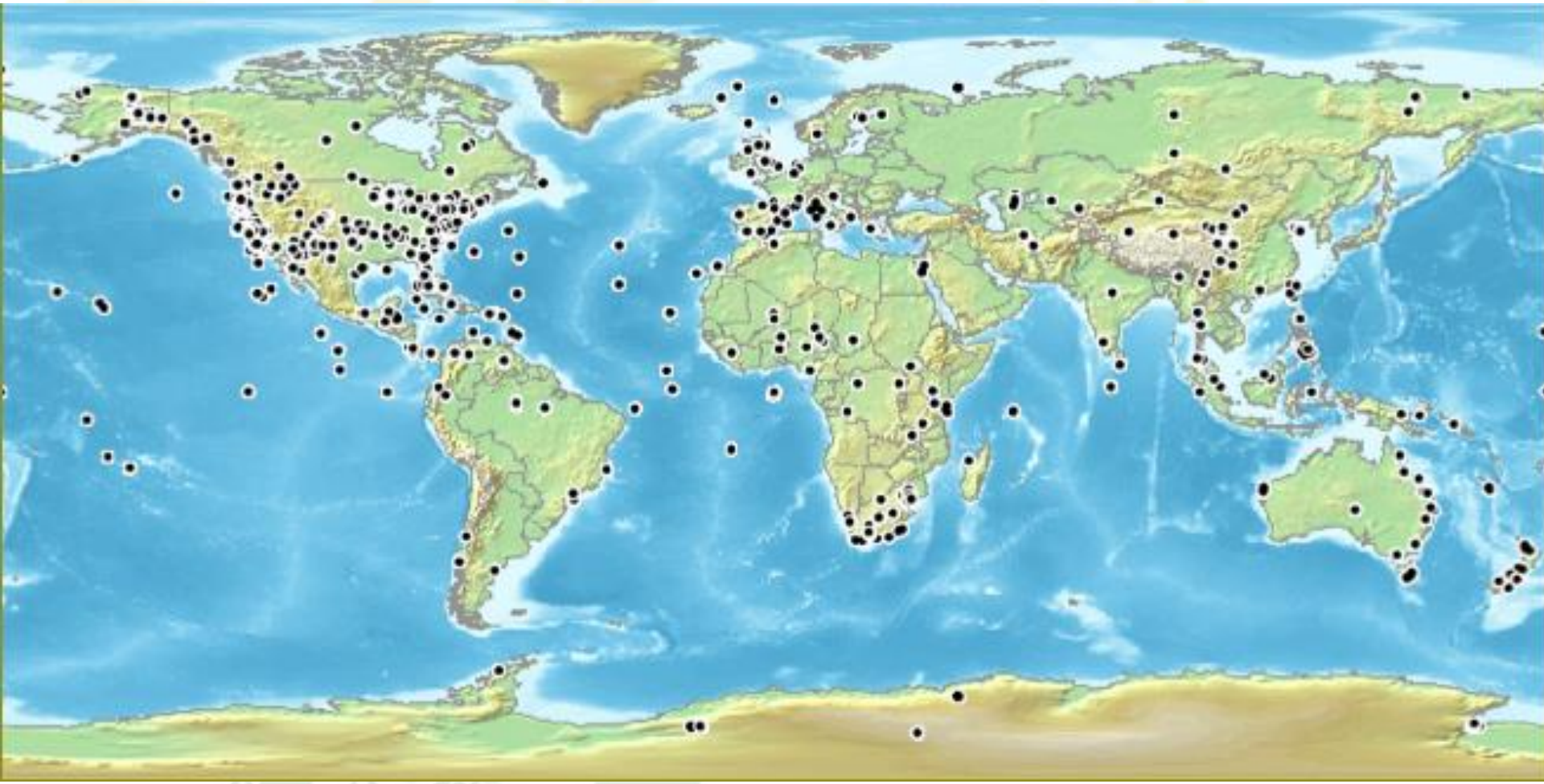
Knowledge Network for Biocomplexity (KNB)
http://knb.ecoinformatics.org/



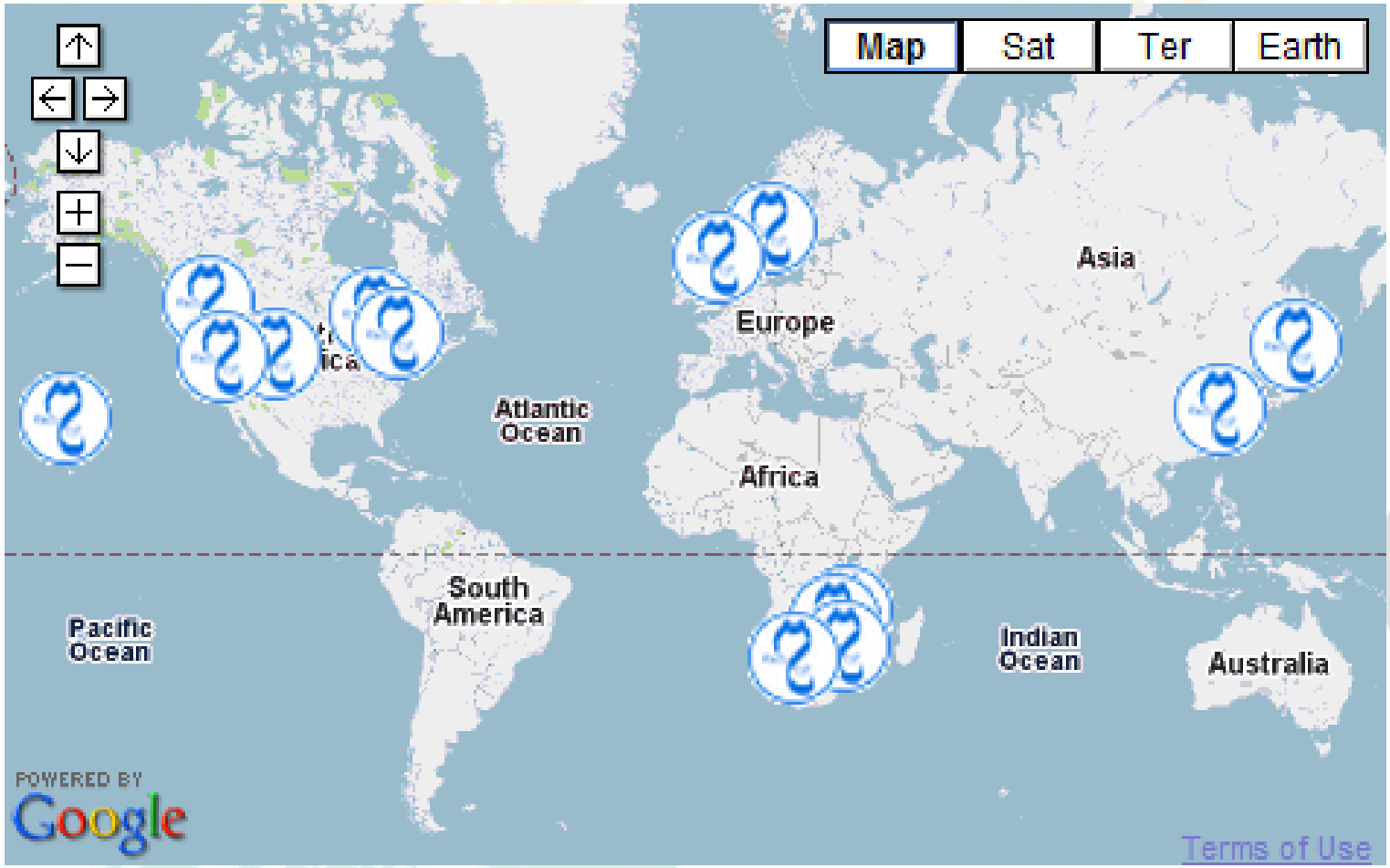
- Morpho: criar, administrar e compartilhar os metadados

- Metacat : armazenamento, busca, recuperação dos dados

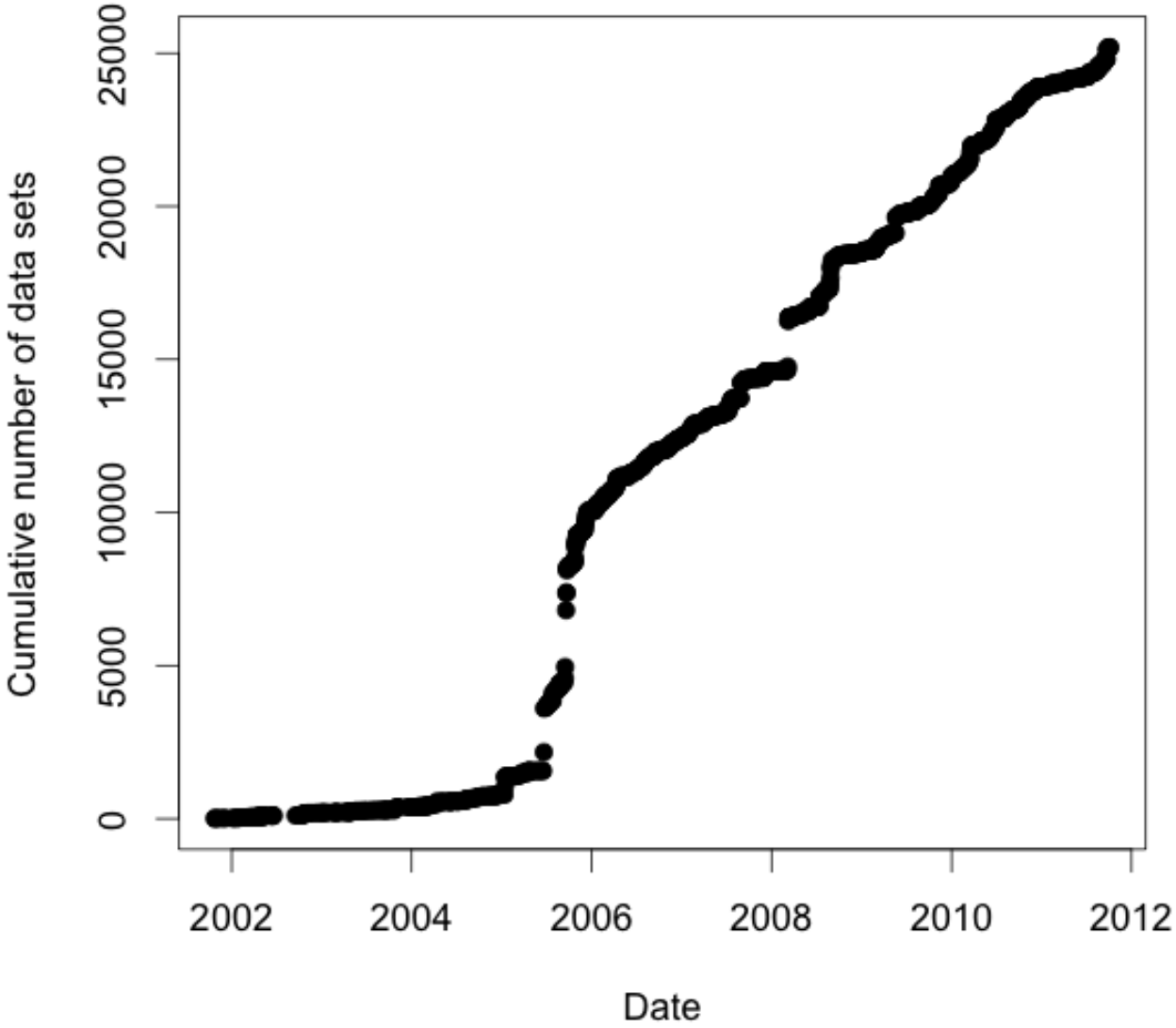
Distribuição de Dados disponíveis no KNB - Knowledge Network for Biocomplexity

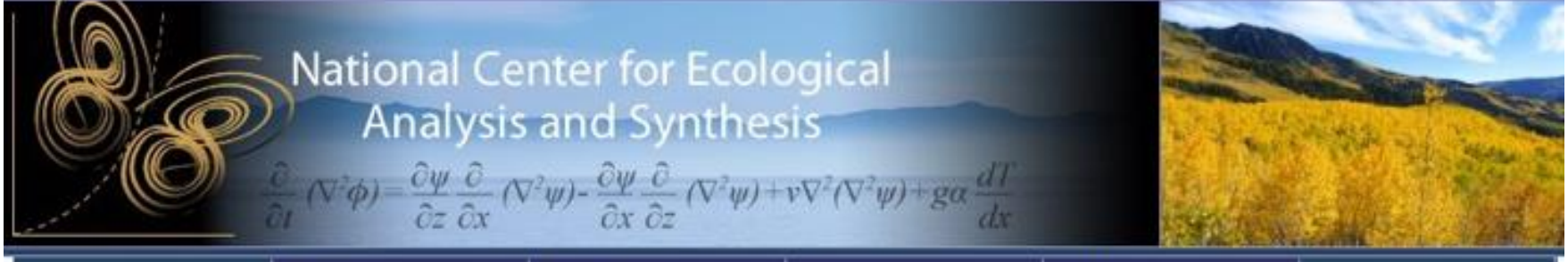


Metacat pelo mundo



Atualmente armazena 25.191 conjuntos de dados, sendo1.296 incluídos em 2011(até agora)





Home | Repository | Register | Login

18 data packages found

Title	Contacts	Organization	Keywords	Actions
» Digital Distribution Maps of the Mammals of the Western Hemisphere ID: connolly.126.2	Brooks Ceballos Luna Ortega Patterson Salazar Sechrest Tognelli Young	NCEAS 4080: Synthesis of the biodiversity knowledge base: Towards a global database of terrestrial vertebrate distributions.	mammals distribution maps Western Hemisphere	View Edit Delete
» Digital Library of Management Laws: California Current Large Marine Ecosystem ID: jekstrom.4.29	Ekstrom	U.C. Santa Barbara	Management Governance Marine law Marine Habitat Marine Environment Monitoring Water Management	View Edit Delete
» Effect of Brassica nigra on the pressure of native consumers on Nassella pulchra in Cheeseboro Canyon ID: nceas.881.3	Orroch	Washington University	Brassica nigra Nassella pulchra rodents	View Edit Delete

Slide: Matt Jones - NCEAS

LTER Home | Intranet

LTER

The US
Long Term Ecological Research
Network

ER Home
gin
arch
owse

Welcome to the LTER Data Catalog

Data are one of the most valuable products of the LTER program. The LTER Network seeks to inform the LTER and broader scientific community by creating well designed and well documented databases and to provide fast, effective, and open access to LTER data via a network-wide information system designed to facilitate data exchange and integration. Currently, the LTER Data Catalog contains entries for over 6000 ecological datasets from 26 LTER Network research sites, and thousands of additional datasets from numerous other ecological field stations and research institutions.

Data Catalog

The LTER Data Catalog includes content from both LTER and non-LTER data sources including PISCO, KNB, etc. By default, search results display only LTER data sources. You may include non-LTER data sources in your search by selecting the check box below.

NEW! When you begin typing in the search box form below, an auto-completion dialog will suggest ranked terms that originate from key-words and titles within the Data Catalog.

Search Term: Advanced Search

Include non-LTER data

Search Reset

LTER Data Policies

The LTER data policy includes three specific sections designed to express shared network policies regarding the release of LTER data products, user registration for accessing data, and the licensing agreements specifying the conditions for data use.

Other Databases

Additional information is available through these value-added data products:

- LTER/USFS Climate / Hydrology Data
- Annual Net Primary Productivity Data

Slide: Matt Jones - NCEAS

://www.jalter.org/index.php?ml_lang=en

JaLTER

HOME SITEMAP

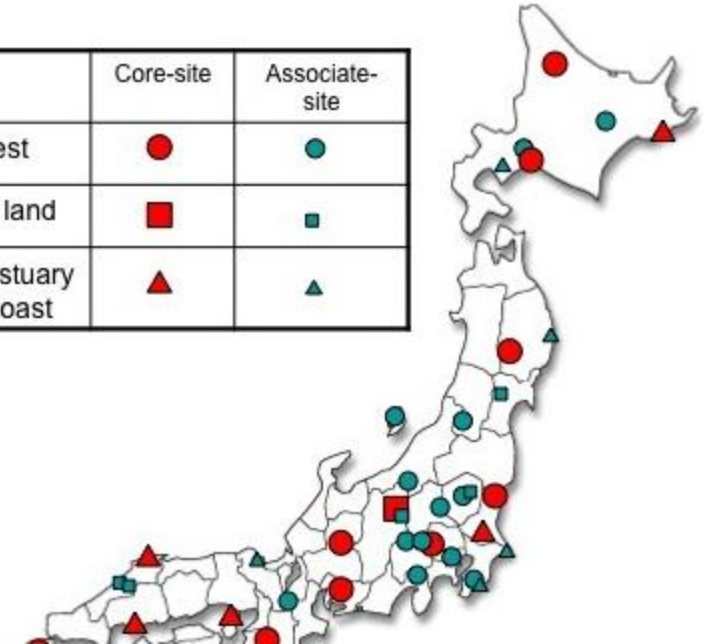
NEWS ABOUT DOWNLOADS LINK CONTACT

JaLTER Metacat Service (Observation Database)

Welcome to JaLTER

Japan Long Term Ecological Research Network.

	Core-site	Associate-site
Forest	●	●
Grass land	■	■
Lake, Estuary and Coast	▲	▲



Language Select

Japanese English

Main Menu

- Home
- News
- About JaLTER
- Locations of Research Sites
- JaLTER Database
- Research Project
- Download
- Contact US
- Links
- Site Map

Search

http://db.cger.nies.go.jp/JaLTER/metacat-en/metacat

JaLTER Data Catalog Search

New search | Japanese skin | JaLTER Data Policy

search for data

All data on this server is "public" . You may search and access to the data catalog without logged-in. Enter a search phrase (e.g. biodiversity) to search for data sets in the data catalog, or simply browse by category using the links below.

Search Data Catalog » advanced search «

Taxonomy

Plant, Invertebrate, Mammal, Bird, Reptile, Amphibian, Fungi, Microbe, Virus

Physical environment

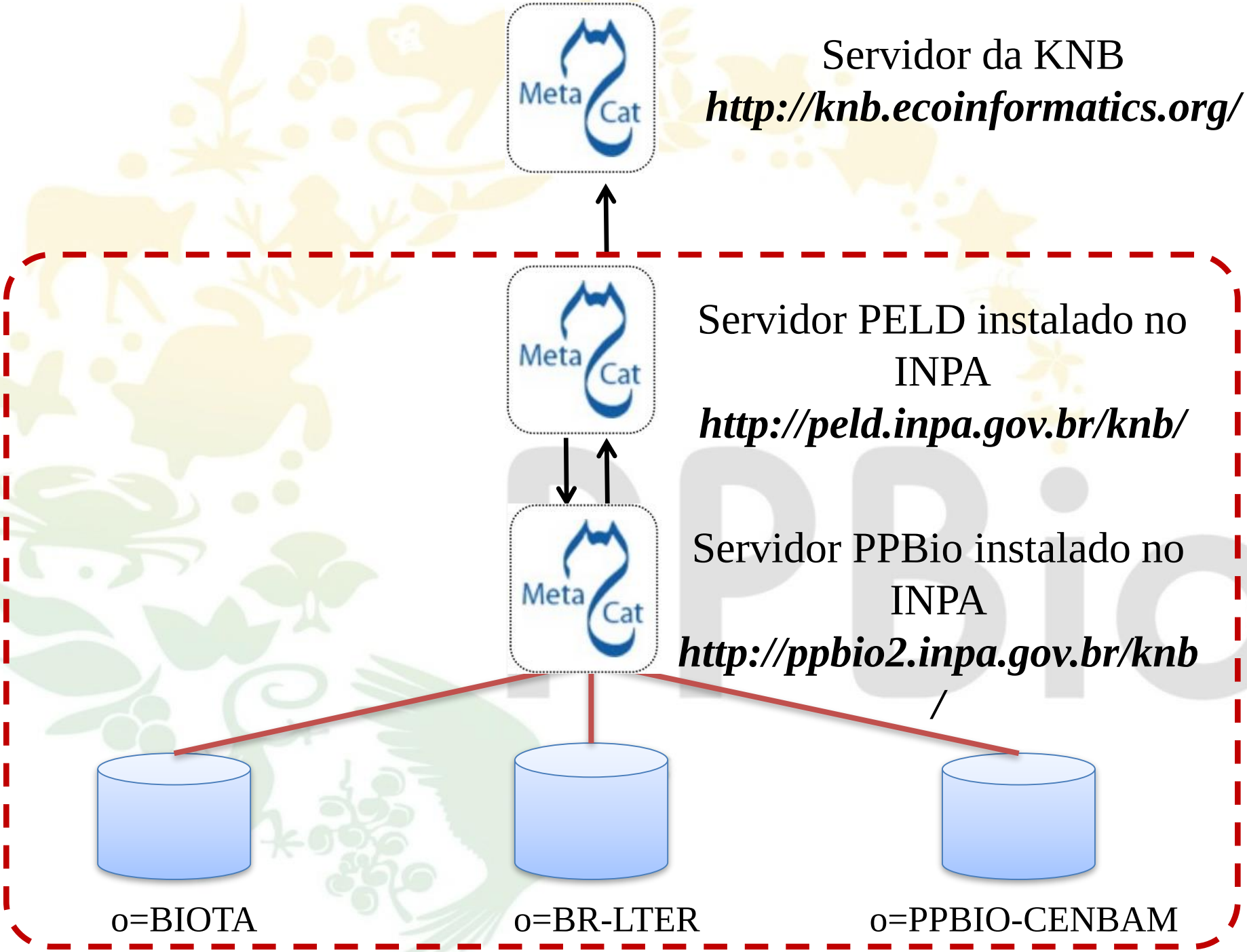
Meteorology, Hydrology, Flux, Satellite

Habitat

Alpine, Aquatic, Beach, Benthic, Desert, Estuary, Forest, Grassland, Marine, Montane, Oceanic, Savanna, Shrubland, Terrestrial, Tundra, Urban, Wetland

22 data packages found

Title	Contacts	Organization	Ke
» Annual water budget in Fukuroyamasawa experimental watershed, Chiba, Japan (1994-1998); 千葉･袋山沢試験流域年水収支データ (1994-1998) ID: JaLTER-Fukuroyama-EM.3.4		Lab. of Forest Hydrology and Erosion Control Engineering, Graduate School of Agricultural and Life Sciences, The University of Tokyo. 東京大学大学院農学生命科学研究科森林科学専攻森林理水及び砂防工学研究室	a 収 co 針 pr w
» Annual water budget in Fukuroyamasawa experimental watershed, Chiba, Japan (1999-2003); 千葉･袋山沢試験流域年水収支データ (1999-2003)		Lab. of Forest Hydrology and Erosion Control	a 収


Servidor da KNB
http://knb.ecoinformatics.org/
Servidor PELD instalado no INPA
http://peld.inpa.gov.br/knb/
Servidor PPBio instalado no INPA
http://ppbio2.inpa.gov.br/knb/
o=BIOTA
o=BR-LTER
o=PPBIO-CENBAM

Ciclo de Vida dos Dados

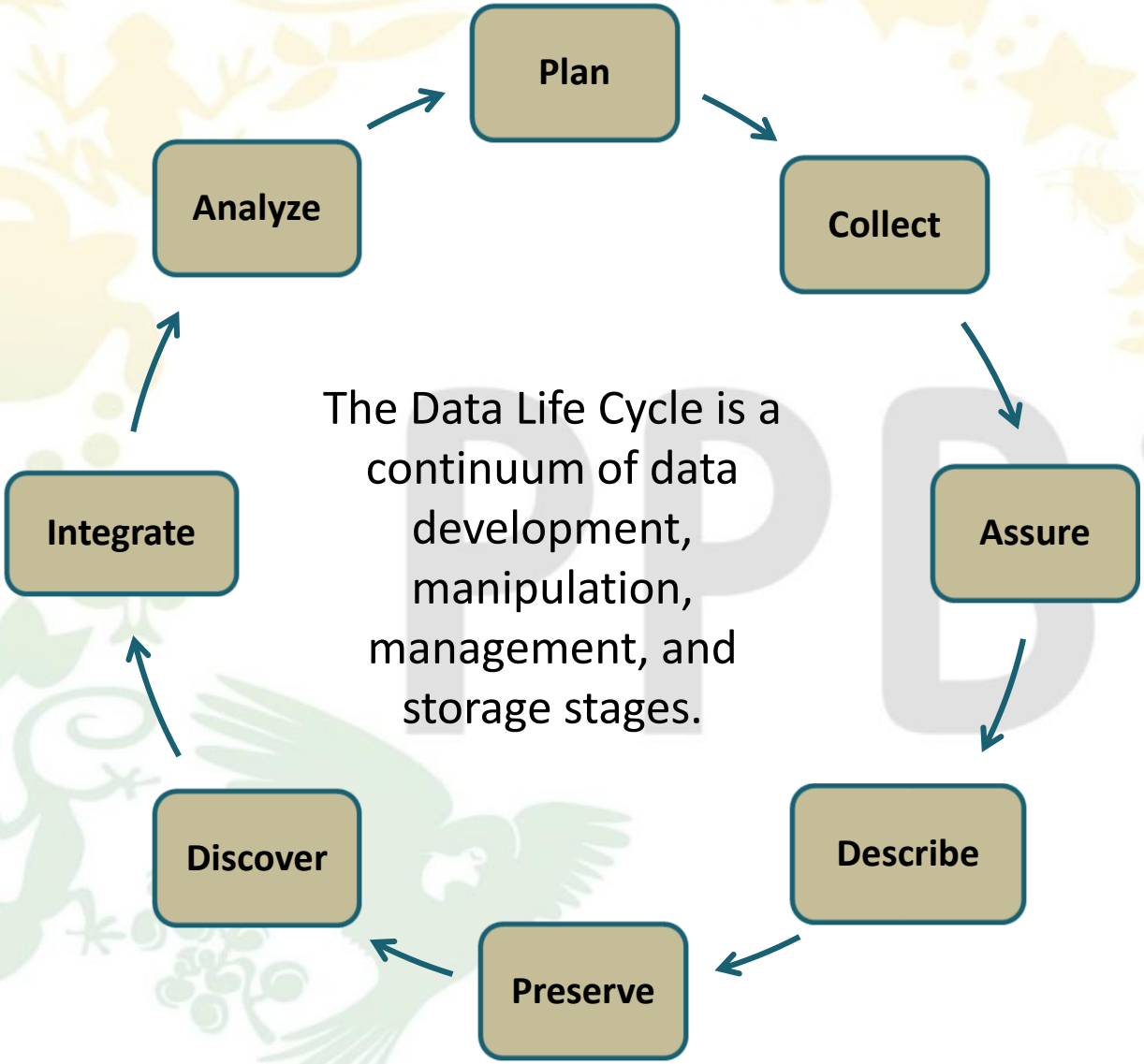


Figura: DataONE - https://www.dataone.org/

Figura: DataONE - https://www.dataone.org/

Antes da Análise de Dados

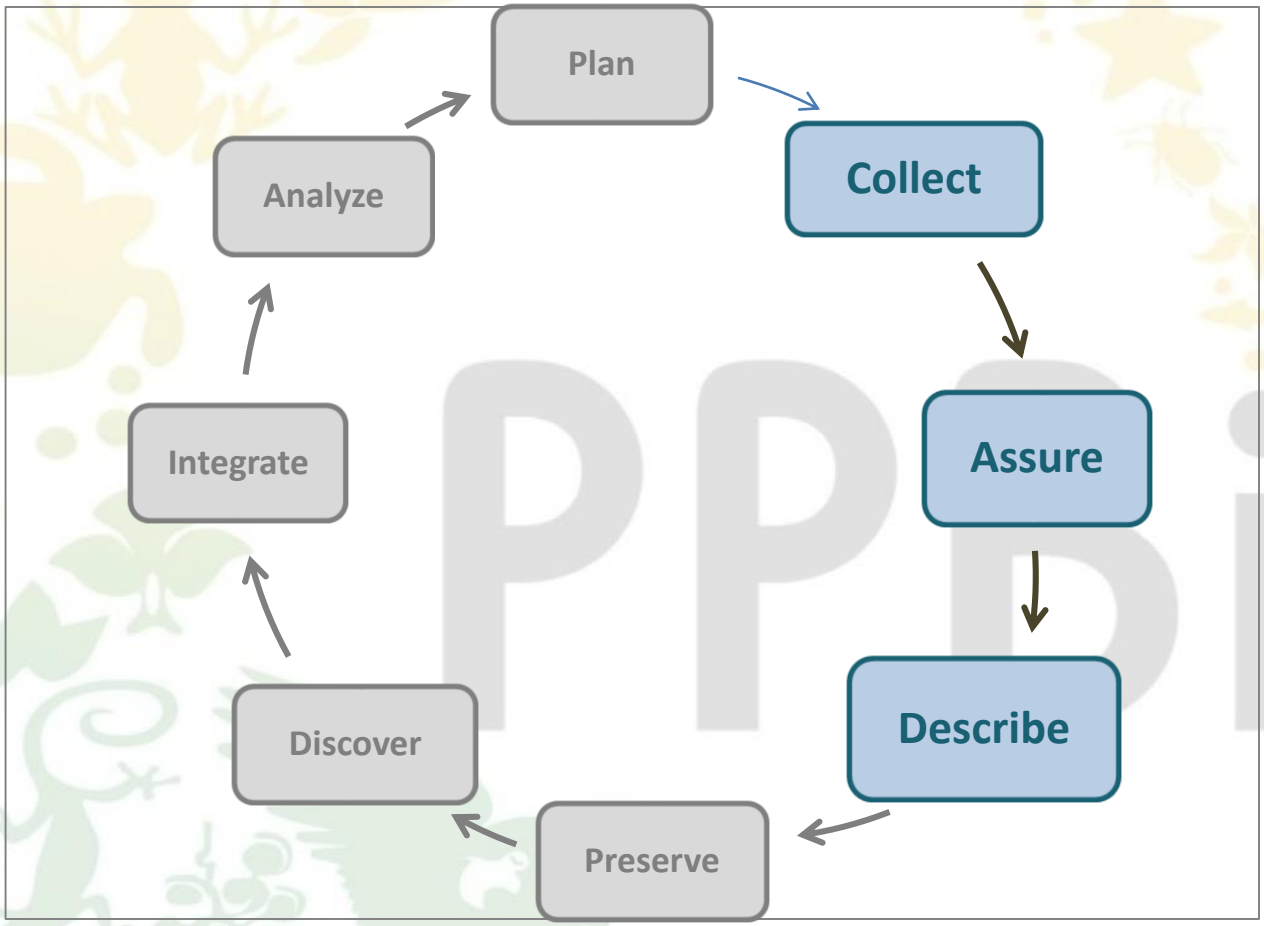


Figura: DataONE - https://www.dataone.org/

Desafios para integração de dados

1	Season	Year	Stream	acefal	aeqpal	amoele	ancist	aphioc	apisto	aucpun	brafre	brycau	brygia	bryinp	calcal	ca
2	Dry1	2005	AC11	0	8	0	0	0	0	0	0	0	6	3	0	
3	Dry1	2005	AC12	0	4	0	0	0	0	0	0	0	0	2	0	
4	Dry1	2005	AC13	0	9	0	0	0	0	0	0	0	0	13	1	
5	Dry1	2005	AC14	0	7	0	0	0	0	0	0	0	0	0	0	
6	Dry1	2005	AC15	0	8	0	0	0	0	0	0	0	0	1	0	
7	Dry1	2005	AC22	0	19	0	0	0	0	0	0	0	3	2	0	
8	Dry1	2005	AC23	0	13	0	0	0	0	0	0	0	39	2	0	
9	Dry1	2005	BO11	0	9	0	0	0	0	0	0	0	6	0	0	
10	Dry1	2005	BO12	0	3	0	1	0	0	0	0	1	26	0	0	
11	Dry1	2005	BO13	0	7	0	0	0	0	0	0	0	0	1	0	
12	Dry1	2005	BO14	0	13	0	0	0	0	0	0	0	6	1	0	
13	Dry1	2005	BO15	0	13	0	0	0	0	0	0	0	0	0	0	
14	Dry1	2005	BO16	0	14	0	0	0	0	0	0	0	0	0	0	
15	Dry1	2005	BO21	0	5	0	0	0	0	0	0	0	35	0	0	
16	Dry1	2005	BO22	0	2	0	1	0	0	0	0	0	45	0	0	
17	Dry1	2005	BR21	0	2	0	1	0	24	0	0	3	0	2	0	
18	Dry1															
19	Dry1															
20	Dry1															

1	parcela	nome_con	narvore	familia	especie	fertil	date	collector	collnumber
2	I5/500		3265				11-Nov-07		
3	I2/3500		17323	Sapindaceae			10-Feb-08		
4	I1/1500		14382	Anacardiaceae	Astronium lecointei Ducke				
5	I1/1500		14485	Anacardiaceae	Tapirira guianensis Aubl.				
6	I2/2500		15673	Anacardiaceae	Spondias mombin L.	sim	2-May-08	Carvalho, F.A	62UAT
7	I2/3500		17556	Anacardiaceae	Tapirira obtusa (Benth.) J.D.Mitch.		11-Feb-08		
8	I2/3500		17567	Anacardiaceae	Thyrsodium spruceanum Salzm. ex Benth.		11-Feb-08		
9	I2/3500		17607	Anacardiaceae	Astronium lecointei Ducke		11-Feb-08		
10	I2/3500		17624	Anacardiaceae	Astronium lecointei Ducke				
11	I2/3500		17637	Anacardiaceae	Tapirira obtusa (Benth.) J.D.Mitch.				
12	I2/3500		17719	Anacardiaceae	Tapirira		9-Feb-08		
13	I2/3500		17790	Anacardiaceae	Astronium		10-Feb-08		
14	I2/3500		17803	Anacardiaceae	Tapirira obtusa (Benth.) J.D.Mitch.		10-Feb-08		
15	I2/500		9123	Anacardiaceae	Tapirira guianensis Aubl.				
16	I2/500		9401	Anacardiaceae	Anacardium spruceanum Benth. ex Engl.				
17	I2/500		9536	Anacardiaceae	Anacardium spruceanum Benth. ex Engl.				
18	I3/1500	Maracatia	6787	Anacardiaceae	Astronium		23-Jul-07		
19	I3/500		5024	Anacardiaceae	Tapirira obtusa (Benth.) J.D.Mitch.	f	8-Nov-07		
20	I3/500		5508	Anacardiaceae	Tapirira obtusa (Benth.) J.D.Mitch.	f	8-Nov-07		

Resultados Comp. 0-5 Monte Cristo [Modo de Compatibilidade] - Microsoft Excel

Início | Inserir | Layout da Página | Fórmulas | Dados | Revisão | Exibição | Suplementos | Acrobat

Colar | Área de T... | Arial | 10 | Fonte | Quebrar Texto Automaticamente | Mesclar e Centralizar | Alinhamento | Geral | Número | Formatação Condicional | Fo como | Est

A8

INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA - INPA

LABORATÓRIO TEMÁTICO DE SOLOS E PLANTAS - LTSP

RESULTADOS DAS ANÁLISES QUÍMICAS DE AMOSTRAS DE SOLO

Projeto: PPBio - Monte Cristo

Amostras Compostas - 0-5 centimetros

Amostra			Prof.	pH	pH	M.O.	N	C	Ca++	Ca++	Mg++	Mg++
LTSP- Nr.	Linha	Km	cm	H2O	KCl	%	%	%	mg/kg	cmolc.kg	mg/kg	cmolc.kg
1	W0	700	0-5	5,62	4,75				76,0	0,38	50,0	0,41
2	W0	700	5-10	5,72	4,59				69,5	0,35	34,0	0,28
3	W0	700	10-20	5,66	4,46				47,5	0,24	17,0	0,14
4	W1	750	0-5	5,67	4,54				68,5	0,34	48,0	0,39
5	W1	750	5-10	5,62	4,42				48,0	0,24	27,0	0,22
6	W1	750	10-20	5,48	4,29				39,0	0,19	15,0	0,12
7	W1	1750	0-5	5,44	4,47				29,0	0,14	39,0	0,32
8	W1	1750	5-10	5,41	4,36				23,5	0,12	30,0	0,25
9	W1	1750	10-20	5,27	4,48				18,0	0,09	22,0	0,18
10	W2	750	0-5	5,3	4,46				33,5	0,17	20,0	0,16
11	W2	750	5-10	5,37	4,38				24,5	0,12	12,0	0,10
12	W2	750	10-20	5,52	4,4				13,0	0,06	10,0	0,08
13	W2	1250	0-5	5,55	4,45				52,0	0,26	39,0	0,32
14	W2	1250	5-10	5,65	4,5				43,0	0,21	31,0	0,26
15	W2	1250	10-20	5,61	4,4				35,0	0,17	22,0	0,18
16	W2	1750	0-5	5,69	4,38				87,5	0,44	117,0	0,96
17	W2	1750	5-10	5,74	4,37				97,0	0,48	136,0	1,12
18	W2	1750	10-20	5,84	4,35				97,5	0,49	139,0	1,14
19	W2	2200	0-5	5,92	4,48				33,5	0,17	35,0	0,29
20	W2	2200	5-10	5,91	4,49				25,0	0,12	23,0	0,19
21	W2	2200	10-20	5,83	4,52				25,0	0,12	21,0	0,17
22	W3	1250	0-5	5,4	4,33				18,0	0,09	19,0	0,16

Ducke Quimica - Microsoft Excel

Início | Inserir | Layout da Página | Fórmulas | Dados | Revisão | Exibição | Suplementos | Acrobat

Obter Dados Externos: Do Access, Da Web, De Texto, De Outras Fontes, Conexões Existentes

Conexões: Atualizar tudo, Conexões, Propriedades, Editar Links

Classificar e Filtrar: Classificar, Filtro, Limpar, Reaplicar, Avançado

Ferramentas de Dados: Texto para colunas, Remover Duplicatas, Validação de Dados, Consolidar, Teste de Hipóteses

R10

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Amostra	linha	km	pH	ca1	ca2	mg1	mg2	k1	k2	al	P	Fe	Zn	Mn
2	1	L1	0	3.58	8.8	0.044	6.1	0.05	20.3	0.052	1.436	2.443	255	0.5	
3	2	L1	1000	3.39	19.8	0.099	17.1	0.141	37.9	0.097	2.801	3.954	258	1	1.4
4	3	L1	2000	3.6	17.8	0.089	20.1	0.165	34.6	0.088	2.42	2.925	234	0.8	1.2
5	4	L1	3000	3.55	15.8	0.079	14.1	0.116	39.4	0.101	2.793	3.472	266	0.9	
6	5	L1	4000	3.72	14.3	0.071	14.6	0.12	31.1	0.08	2.197	3.311	274	0.7	0.6
7	6	L1	5500	3.84	8.8	0.044	7.6	0.063	24.1	0.062	1.696	3.826	282.5	0.5	0.2
8	7	L1	6500	4.65	5.3	0.026	4.6	0.038	26	0.066	0.881	4.018	373.5	0.2	
9	8	L1	7500	3.62	8.3	0.041	11.1	0.091	26.1	0.067	0.93	3.954	147.5	0.3	
10	9	L2	0	4.09	23.8	0.119	27.1	0.223	36.2	0.093	0.523	9.548	47	0.6	
11	11	L2	1000	3.83	12.3	0.061	12.1	0.1	39.2	0.1	2.208	4.533	291	0.3	0.6
12	13	L2	2000	4.02	8.3	0.041	7.6	0.063	23	0.059	1.431	2.54	294	1.2	0.1
13	15	L2	3000	3.92	10.8	0.054	9.6	0.079	25.1	0.064	1.357	2.829	208.5	1	
14	17	L2	4000	3.92	17.3	0.086	9.1	0.075	22.7	0.058	1.008	5.562	178	0.7	
15	19	L2	5500	3.48	28.8	0.144	23.1	0.19	47.8	0.122	2.886	4.276	208	1.5	3.3
16	20	L2	6500	3.79	19.3	0.096	12.6	0.104	27	0.069	0.864	4.597	155	0.6	
17	21	L2	7500	3.92	4.8	0.024	3.1	0.026	10.8	0.028	0.516	3.793	49	0.4	
18	23	L3	500	3.9	7.8	0.039	7.6	0.063	21.7	0.055	1.476	1.704	335.5	0.5	
19	25	L3	1500	4.82	0	0	0.1	0.001	8.4	0.021	0.318	0.997	49	0.2	
20	27	L3	2500	5.05	3.8	0.019	2.1	0.017	10	0.026	0.082	1.222	13.5	0.7	
21	29	L3	3500	4.98	24.3	0.121	17.6	0.145	34.2	0.087	0.392	3.15	21.5	0.7	
22	31	L3	4500	4.75	2.8	0.014	0.6	0.005	8.5	0.022	0.092	0.772	15.5	0.3	
23	32	L3	5500	4.91	12.3	0.061	12.1	0.1	25.3	0.065	1.092	1.736	580	0.9	
24	33	L3	6500	4.46	7.3	0.036	7.1	0.058	17.2	0.044	0.618	3.086	24	0.8	
25	34	L3	7500	3.75	1.8	0.009	2.1	0.017	8.4	0.021	0.552	2.572	103	0.3	
26	35	L4	500	4.46	12.8	0.064	5.6	0.046	12.3	0.031	0.023	1.511	12	0.3	
27	36	L4	1500	4.49	3.8	0.019	1.6	0.013	13.3	0.034	0.252	1.222	47	0.3	

Ducke Quimica

Pronto

Chaves Primárias

- Precisam ser suficientemente detalhadas para que os dados possam ser integrados

TABELA 13: DADOS ESCARABEÍDEOS

Concessão	Módulo	Trilha	Parcela	Segmento	Campanha	Levantamento	Armadilha	Lote

TABELA 15: DADOS ÁRVORES

Concessão	Módulo	Trilha	Parcela	Segmento	Campanha	Levantamento	Número do Indivíduo	Espécie	DAP	Número da coleta	Número de duplicatas

TABELA 9: DADOS PRIMATAS

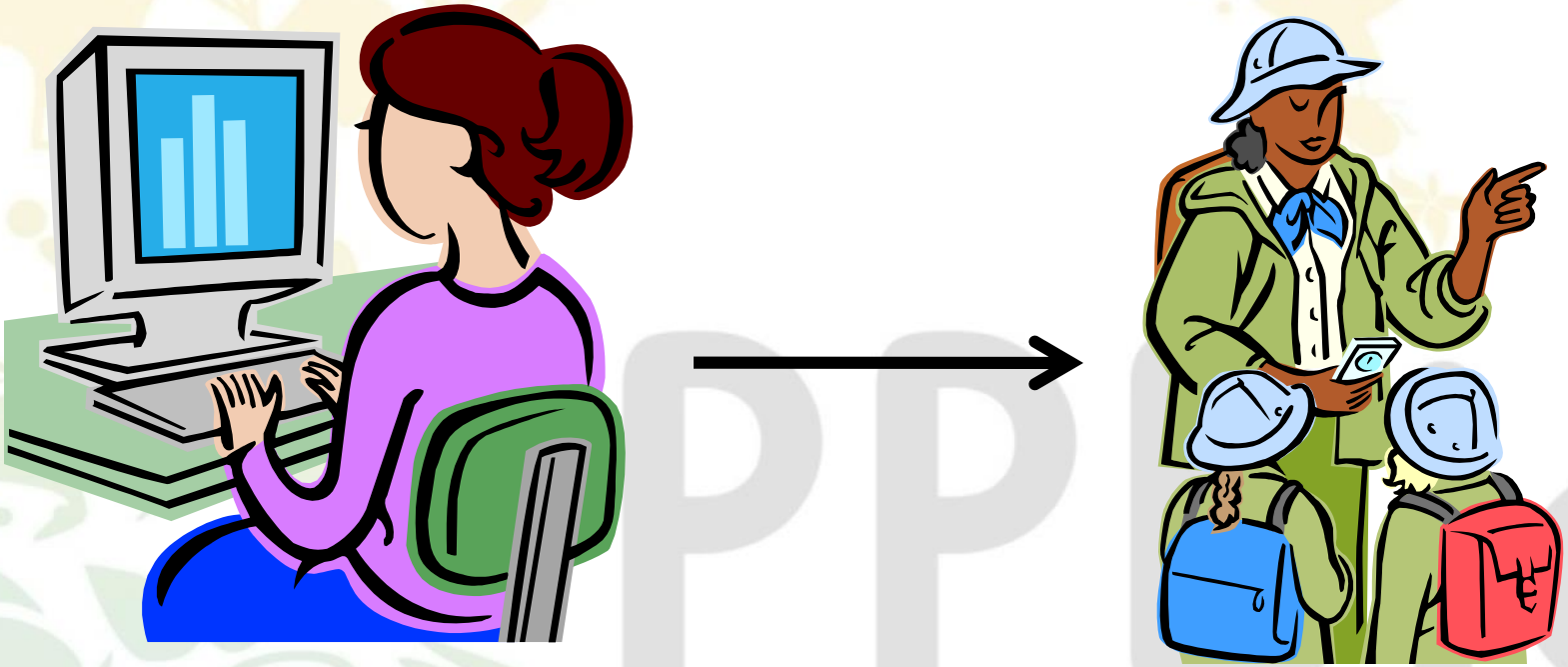
Concessão	Módulo	Trilha	Campanha	Levantamento	Temp.	Hora	Espécie	Num. no grupo	Dist. na trilha	Dist. da trilha	

http://ppbio.inpa.gov.br/Port/dadosinvent/

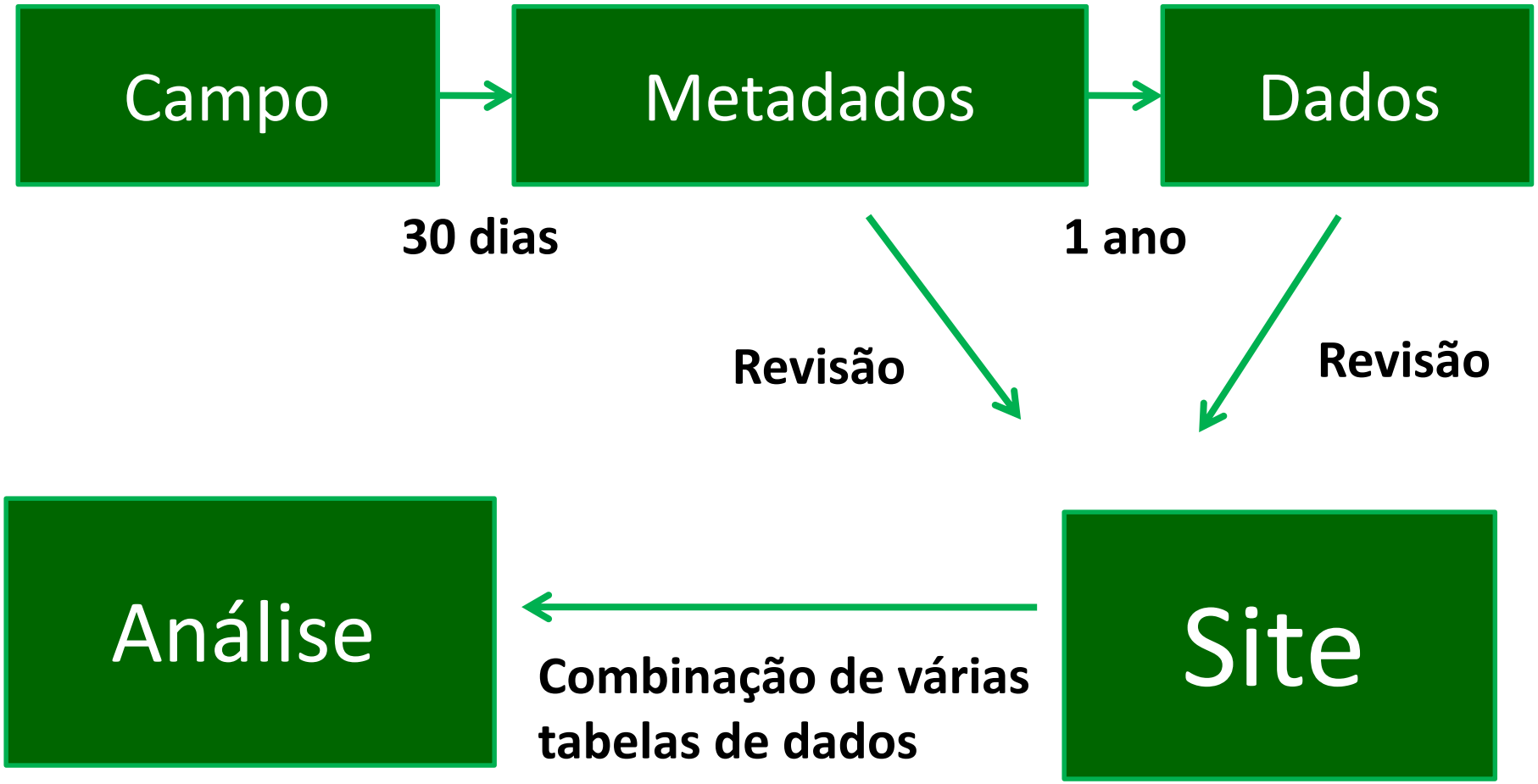
TABELAS RELACIONADAS:

- Pessoas;
- Campanhas;
- Levantamentos;
- Dados de Campo;
- Sugestão de planilha de campo

O “filtro humano”



Como funciona hoje



Para um repositório de dados funcionar, é necessário:

I. “Equipamentos”

Pessoas dedicadas a receber e checar os dados e metadados, dialogar com os geradores dos dados para corrigir os erros e finalmente disponibilizar os dados e metadados on line

Servidores para armazenar os dados e metadados

Para um repositório de dados funcionar, é necessário:

II. Decisões

- Arquitetura da rede de recebimento e disponibilização dos dados (recomenda-se uma rede com um ponto central, mas vários pontos regionais, interligados)
- Variáveis mínimas para as chaves primárias das tabelas de dados (descritores de tempo, espaço, e responsáveis pela amostragem, p.ex)
- Adoção de algum(s) modelo de metadados

Para um repositório de dados funcionar, é necessário:

III. Capacitação

- Das **pessoas** dedicadas a gerir os dados e metadados, para serem capazes de detectar os erros e dialogar com os geradores dos dados
- Dos **fornecedores de dados** (pesquisadores e afins), para entender o que são e como escrever metadados, e para construir as tabelas de dados de forma que as informações originais não sejam sumarizadas e que as chaves primárias importantes estejam presentes

OBRIGADA!

http://ppbio.inpa.gov.br

flaviapezzini@gmail.com