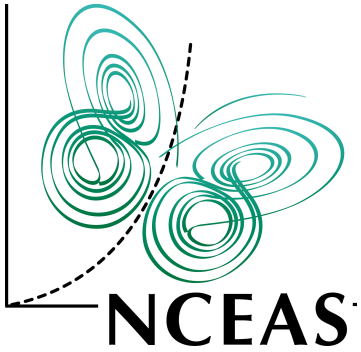
Data Management for Synthesis

Ben Leinfelder
Matthew B. Jones

National Center for Ecological Analysis and Synthesis (NCEAS)
University of California Santa Barbara

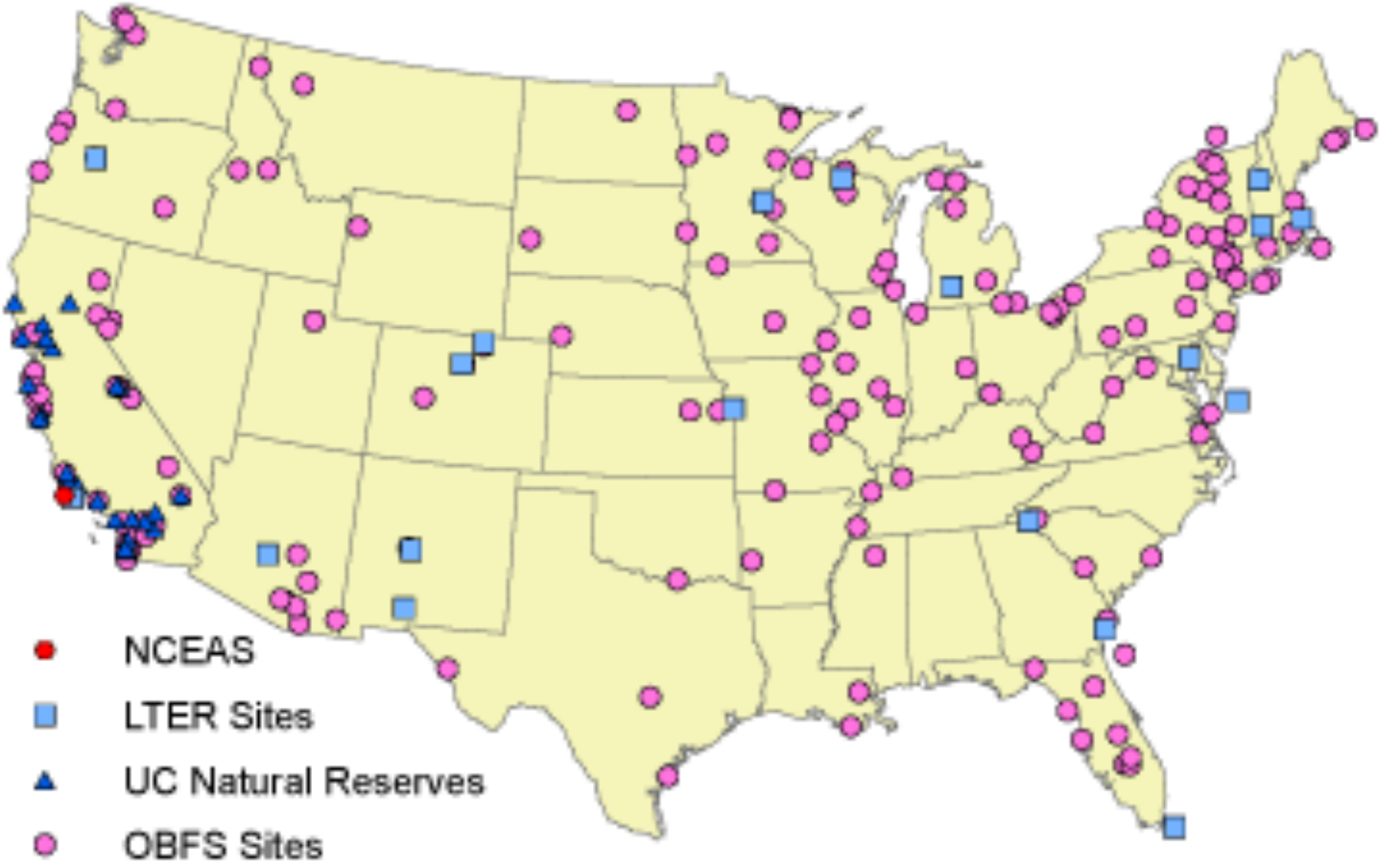


LNCC GBIF workshop
August 25, 2014

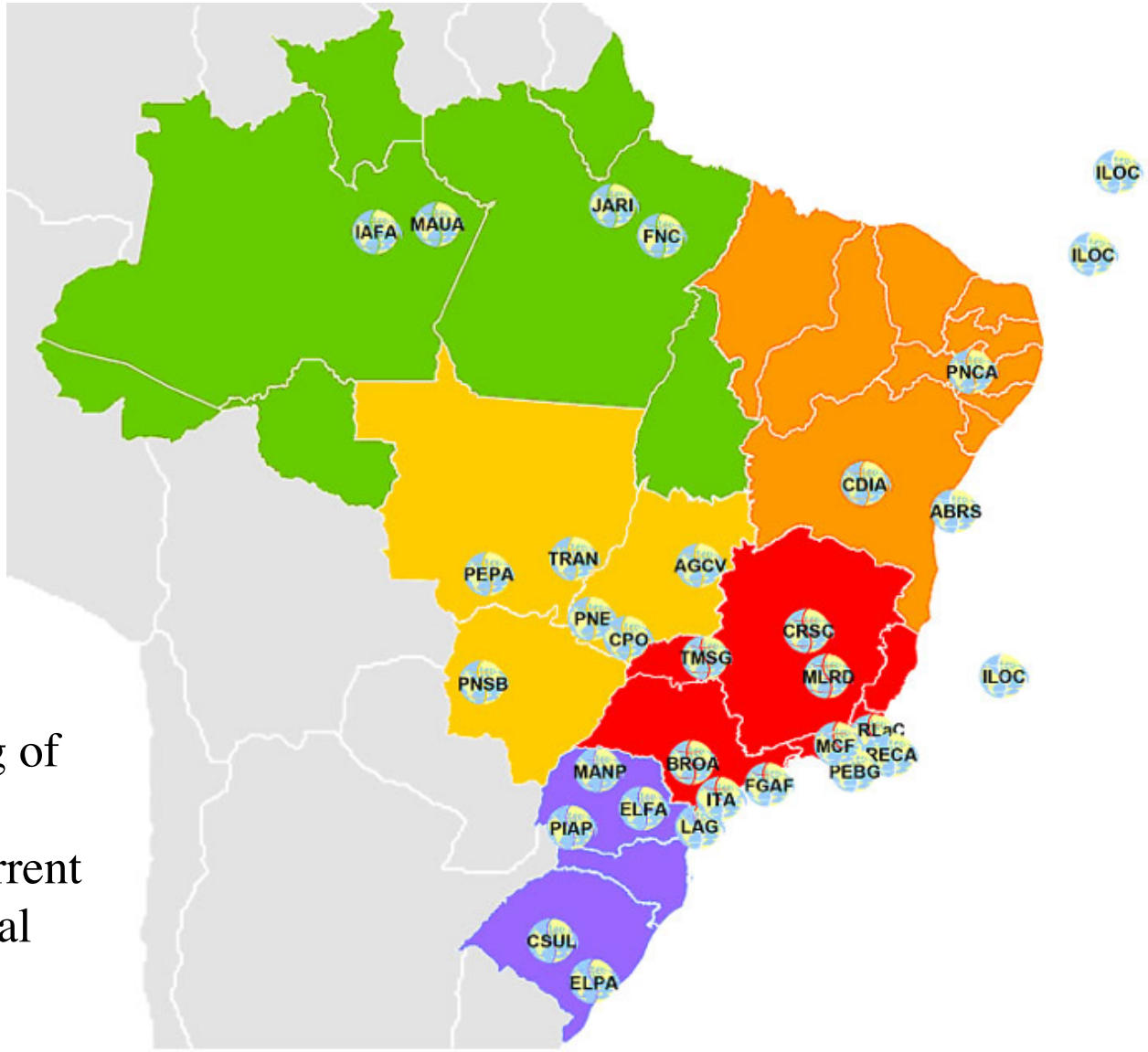
Barriers to Synthesis

- Data not preserved
  - Tiny proportion of ecological data are readily available
- Dispersed, isolated repositories
  - Each community has its own; disconnected; underutilized
- Lack of software interoperability
  - Metacat, DSpace, Mercury, iRODS, XMCat, OPeNDAP, ...
- Heterogeneous data
  - Many data formats, metadata formats, and varying semantics

Dispersed data from field stations



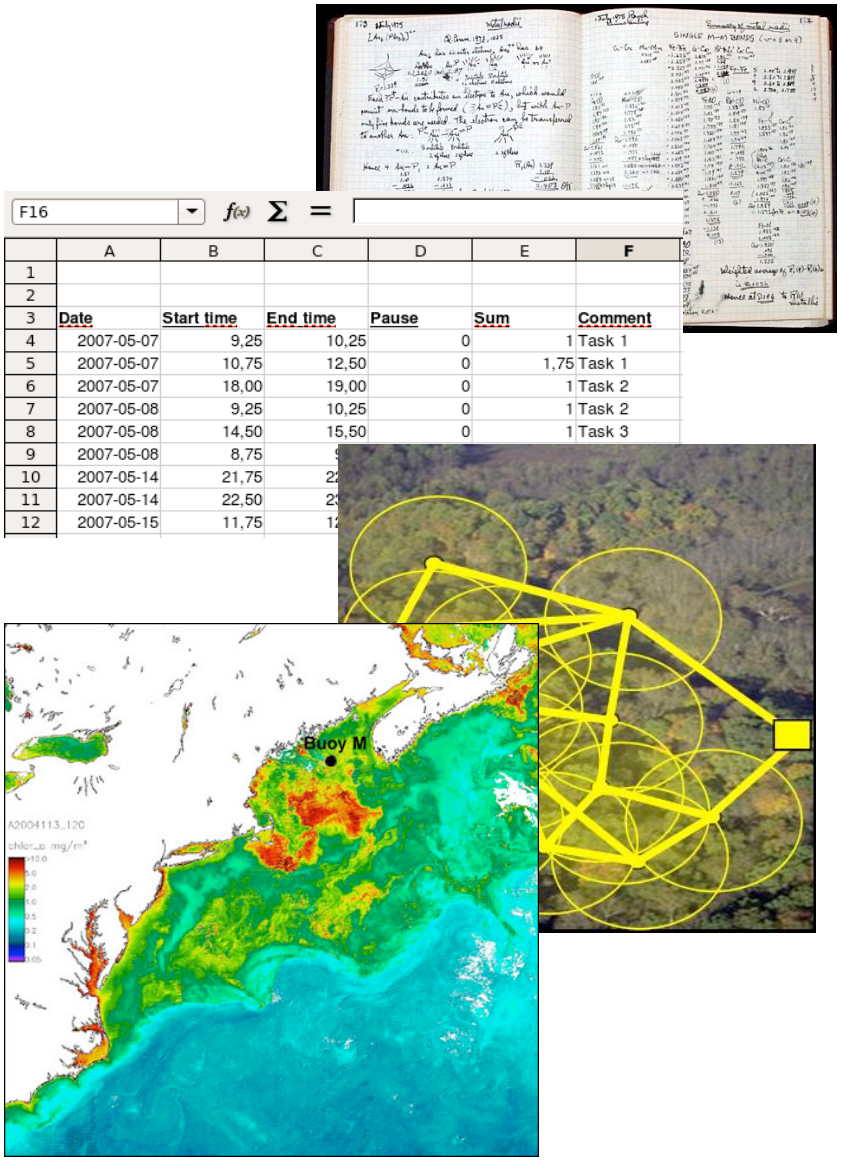
PELD Sites



ILTER:
"increase understanding of global ecosystems and propose solutions to current and future environmental problems"

What data are in scope?

- Biological
  - e.g., Ecosystem, Organism, Population, Species, Community, Biome, Gene
- Environmental
  - e.g., Atmospheric, Chemical, Ecological, Hydrological, Oceanographic, Physical
- Social
  - e.g., Land use, human population
- Economic
  - e.g., trade, ecosystem services, resource extraction



	A	B	C	D	E	F
1						
2						
3	Date	Start time	End time	Pause	Sum	Comment
4	2007-05-07	9,25	10,25	0	1	Task 1
5	2007-05-07	10,75	12,50	0	1,75	Task 1
6	2007-05-07	18,00	19,00	0	1	Task 2
7	2007-05-08	9,25	10,25	0	1	Task 2
8	2007-05-08	14,50	15,50	0	1	Task 3
9	2007-05-08	8,75				
10	2007-05-14	21,75				
11	2007-05-14	22,50				
12	2007-05-15	11,75				

Metadata and data heterogeneity

- Every community has
  - many data schemas
    - one for each project and person
  - many data formats
    - ASCII, NetCDF, HDF, GeoTiff, ...
  - many metadata schemas
    - Biological Data Profile, Darwin Core, Dublin Core, Ecological Metadata Language, Open GIS schemas, ISO Schemas, ...
- Accepting this heterogeneity is critical

Biodiversity data heterogeneity

Space **Time** **Taxa**

(a) Georgia Coastal Ecosystems LTER (EML)

Site	Community	Treatment	Replicate	Species_Code	ITIS_TSN	Plant_Mass	Plant_Mass_m2
GCE6	BorJun	N	1	JROM	39238	75.13	300.5
GCE6	BorJun	N	2	JROM	39238	179.81	719.2
GCE6	BorJun	N	3	JROM	39238	443.20	1772.8
...	...	...	...	...	...	...	...

(b) *Mephitis mephitis* specimen record (DarwinCore)

Cat	Code	Year	Collector	Long	Lat	Cat	Sci_Name	Institution	Species
4815	Mamm	1947	DOWNEY,J.	-109.34	37.86	null	Mephitis mephitis	ESTOR UMNH	mephitis
6767	Mamm	1948	NEILSON,L.	-109.52	40.41	null	Mephitis mephitis	ESTOR UMNH	mephitis
6768	Mamm	1948	NEILSON,L.	-109.52	40.41	null	Mephitis mephitis	ESTOR UMNH	mephitis
...	...	...	...	...	...	...	...	...	...

(c) NOAA Ocean Buoy Data (Station 46069 - South Santa Rosa Island, CA)

YYYY	MM	DD	hh	mm	WD	WSPD	GST	WVHT	DPD	APD	MWD	BAR	ATMP	WTMP	DEWP	VIS	TIDE
2006	01	01	00	00	284	6.9	8.2	3.50	13.79	8.11	275	1013.2	15.5	14.3	999.0	99.0	99.00
2006	01	01	01	00	287	6.0	7.6	3.13	13.79	8.08	271	1013.5	15.3	14.3	999.0	99.0	99.00
2006	01	01	02	00	276	4.3	5.6	3.20	13.79	8.49	272	1013.7	14.9	14.4	999.0	99.0	99.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

(d) Macroecological data for fossil occurrences (Paleobiology Database)

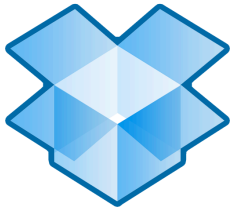
coll_no	genus_name	occ_no	locomot	life_habit	diet1	composition	10mybin	
1	Australosutura	1	mobile	epifaunal	detritivore	chitin	Carboniferous	1
1	Carbonocoryphe	2	mobile	epifaunal	detritivore	chitin	Carboniferous	1
1	Thigriffides	3	mobile	epifaunal	detritivore	chitin	Carboniferous	1
...	...	...	...	...	...	...	...	

Synthesis requires sharing data

- NCEAS' approach to data sharing
  - Deal with data heterogeneity
  - Distributed data management
  - Centralized search
  - Semi-automated analysis tools

- A grass-roots network with global partners
  - NCEAS, LTER, iLTER, PISCO, ESA, SanParks, SAEON, TERN, TEAM, ...

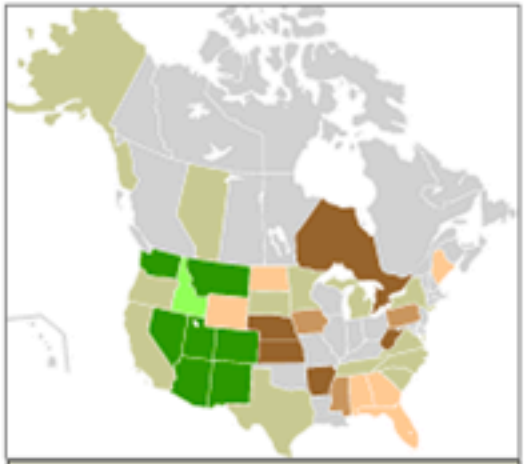
Software diversity



VEGBANK

Find Plots

Browse plots
Simple search
Search with a map
Advanced plot search



Map Key: plots Larger Map		
1–49	50–99	100–249
250–999	1,000–3,000	> 3,000

Recently Added Plots

Project (view all)	Added
FMNE – NatureServe	11-Sep-

Plant Taxa

What is a plant concept?
Browse plants
Search plants
Submit plants

Plant Communities

What is a community?
Search communities
Submit communities

Supplemental Data

People
Stratum methods
Cover methods
Projects
References
Search supplemental data

Data in VegBank

Plots	72,857
--Classified Plots	60,213
----to NVC communities	5,478

News

» **Map plots:** Example | Datacart | Multiple Datasets (Requires Login)
» Save Your Datacart | Edit Datasets
» Create a Constancy Table

My VegBank Account

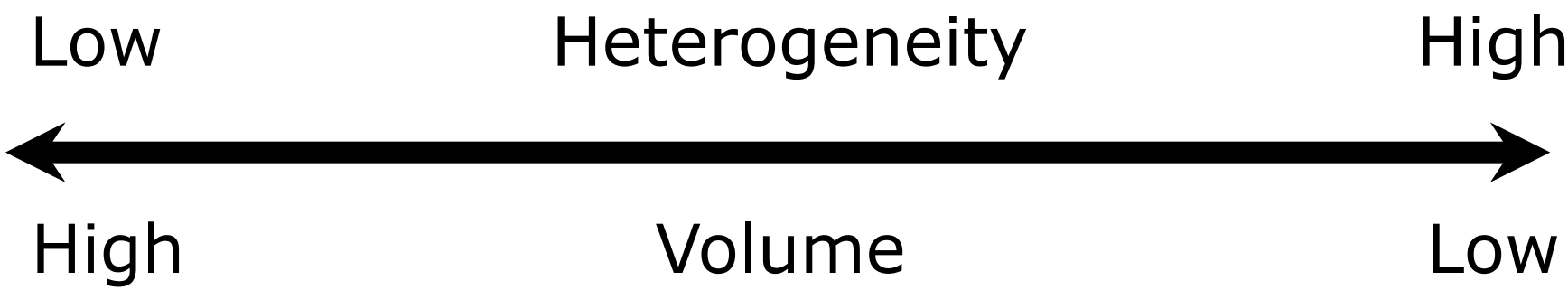
Edit profile information
Manage datasets

Learn About VegBank

What is VegBank?
What is a plot?
FAQ
Tutorial
Cite or link to VegBank
Terms of use
Site map
Contact

Contribute Plot Data

Data Heterogeneity



- Tight coupling
- Simple subsetting
- Explicit semantics

- Loose coupling
- Hard subsetting
- Limited semantics

VEGBANK

knb

MetaCat

Solutions

- Preserve data
- Adopt standards

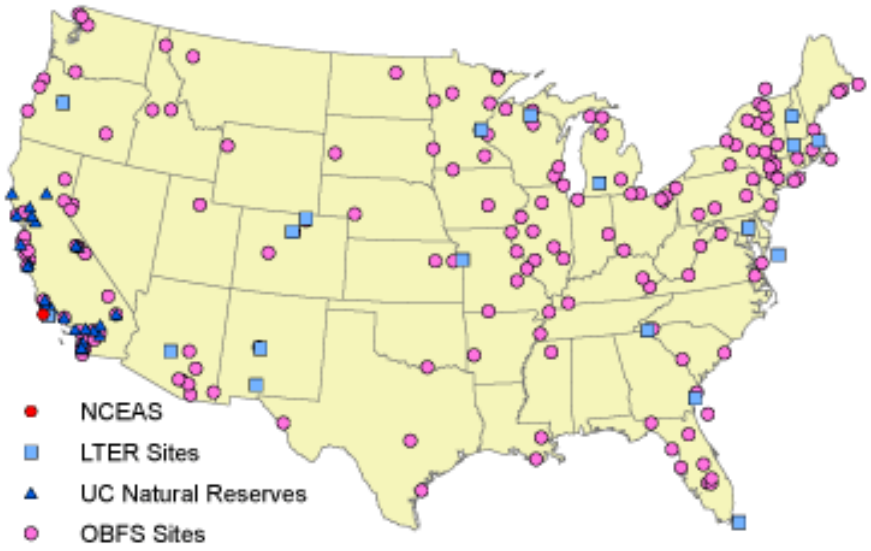
<EML/>

- Create networks
- Create interoperable software

NCEAS

PRESERVE DATA

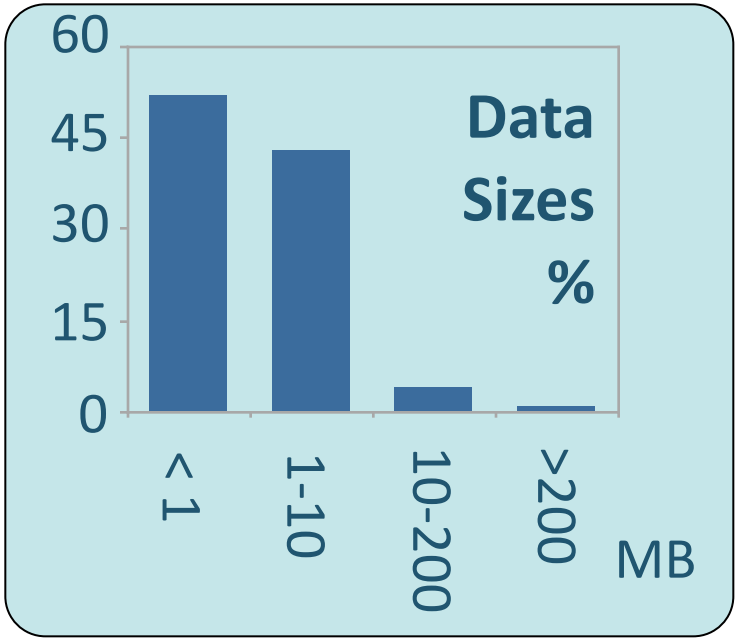
Preserve data in the KNB



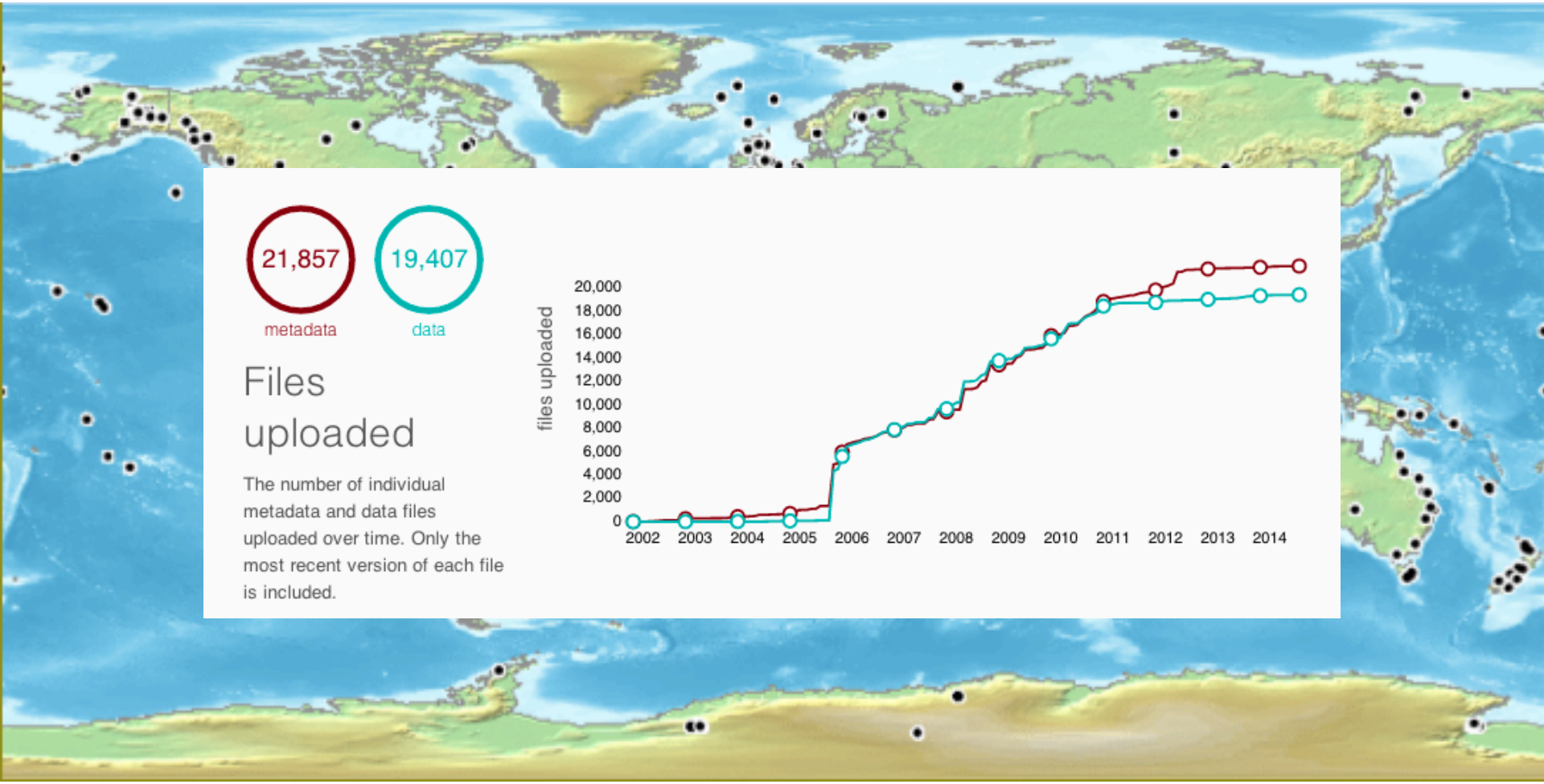
Data Types

- Ecological
- Environmental
- Demographic
- Social/Legal/Economic

- **Diverse Contributors**
  - Individual investigators
  - Field stations and networks
  - Government agencies
  - Non-profit partnerships
  - Scientific Societies
  - Synthesis centers



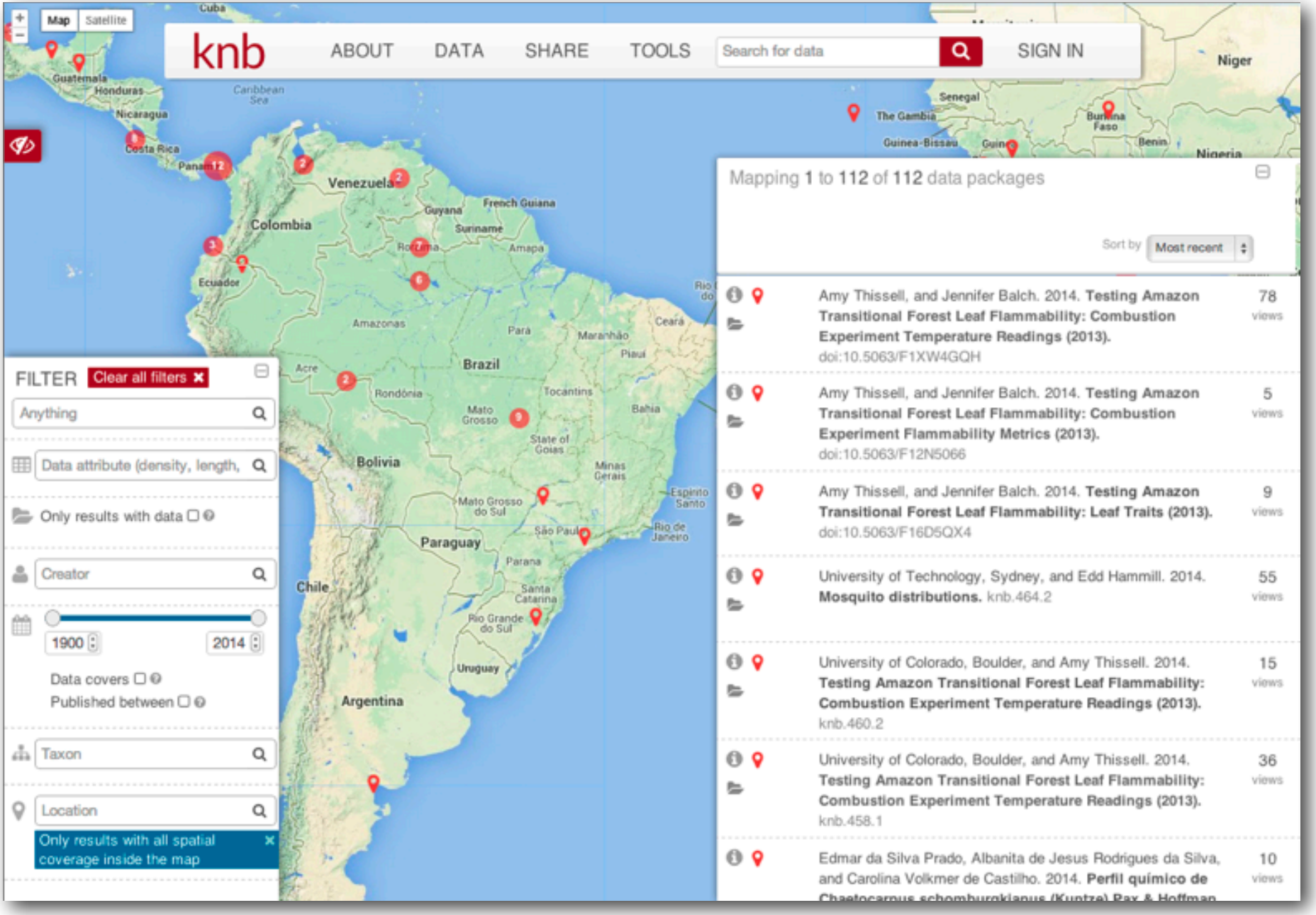
Knowledge Network for Biocomplexity Data Distribution



Metacat Data Server

- Data and metadata management
- Store, search, and document data
- Customizable web-based search interface
- Web metadata entry tool
- DOI Support

- Runs on Linux, Windows, MacOS
- Replication capabilities
- Postgres or Oracle backend
- OAI-PMH harvester
- GPL open source license

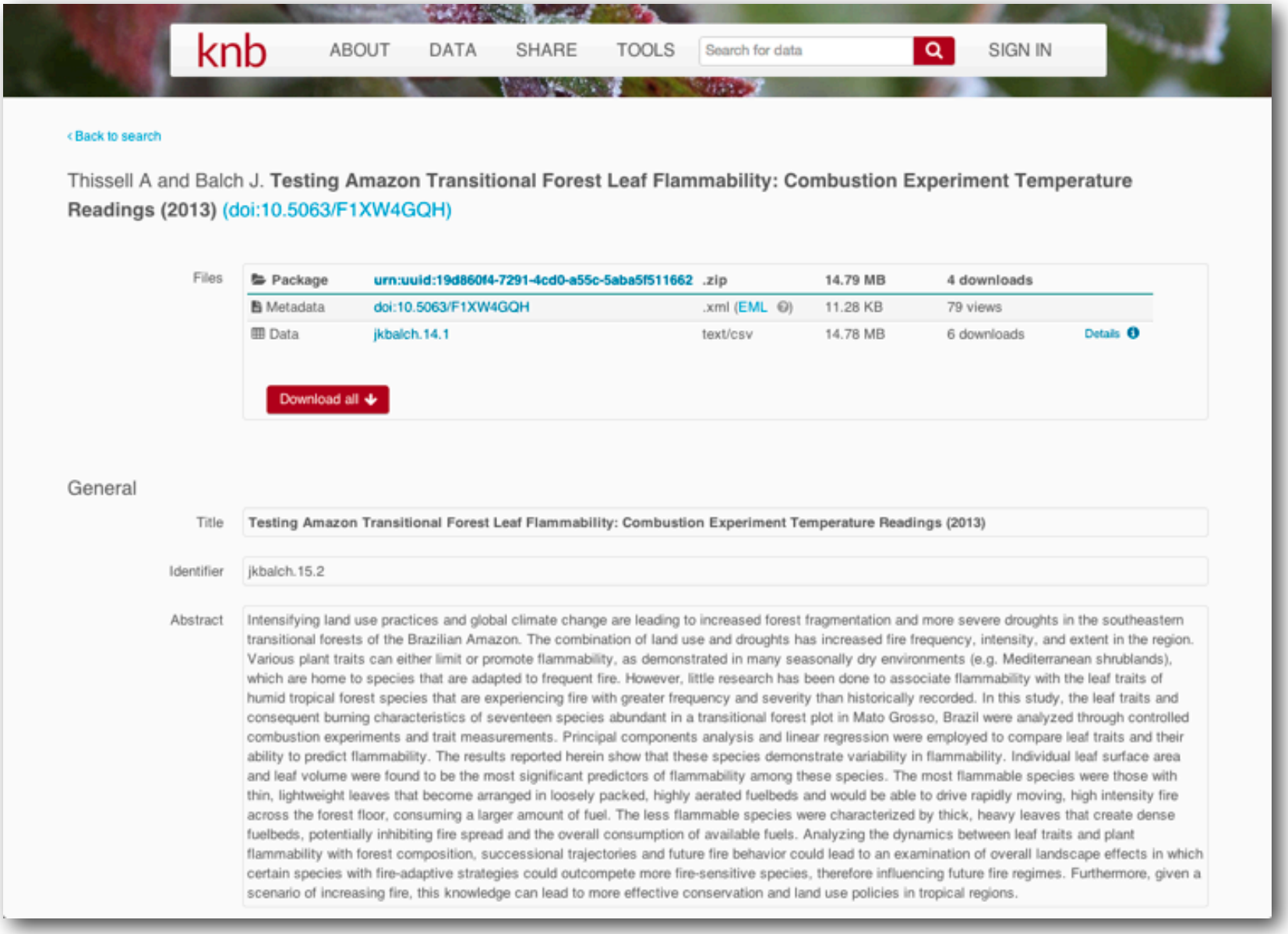

knb
ABOUT
DATA
SHARE
TOOLS
Search for data
SIGN IN
Map
Satellite
Mapping 1 to 112 of 112 data packages
Sort by
Most recent
FILTER
Clear all filters
Anything
Data attribute (density, length,
Only results with data
Creator
1900
2014
Data covers
Published between
Taxon
Location
Only results with all spatial coverage inside the map
Amy Thissell, and Jennifer Balch. 2014. Testing Amazon Transitional Forest Leaf Flammability: Combustion Experiment Temperature Readings (2013).
doi:10.5063/F1XW4GQH
78 views
Amy Thissell, and Jennifer Balch. 2014. Testing Amazon Transitional Forest Leaf Flammability: Combustion Experiment Flammability Metrics (2013).
doi:10.5063/F12N5066
5 views
Amy Thissell, and Jennifer Balch. 2014. Testing Amazon Transitional Forest Leaf Flammability: Leaf Traits (2013).
doi:10.5063/F16D5QX4
9 views
University of Technology, Sydney, and Edd Hammill. 2014. Mosquito distributions. knb.464.2
55 views
University of Colorado, Boulder, and Amy Thissell. 2014. Testing Amazon Transitional Forest Leaf Flammability: Combustion Experiment Temperature Readings (2013).
knb.460.2
15 views
University of Colorado, Boulder, and Amy Thissell. 2014. Testing Amazon Transitional Forest Leaf Flammability: Combustion Experiment Temperature Readings (2013).
knb.458.1
36 views
Edmar da Silva Prado, Albanita de Jesus Rodrigues da Silva, and Carolina Volkmer de Castilho. 2014. Perfil químico de
10 views
Brazil
Bolivia
Paraguay
Chile
Argentina
Colombia
Venezuela
Ecuador
Peru
Guyana
Suriname
French Guiana
Amazonas
Para
Maranhão
Ceará
Piaui
Tocantins
Bahia
Mato Grosso
State of Goias
Minas Gerais
Espirito Santo
Rio de Janeiro
São Paulo
Mato Grosso do Sul
Parana
Santa Catarina
Rio Grande do Sul
Uruguay
Rondônia
Acre
Amapa
Roraima
Cuba
Guatemala
Honduras
Nicaragua
Costa Rica
Panama
Caribbean Sea
Senegal
The Gambia
Guinea-Bissau
Burkina Faso
Benin
Niger
Nigeria

ADOPT STANDARDS

Metadata and data heterogeneity

- Every community has
  - many data schemas
    - one for each project and person
  - many data formats
    - ASCII, NetCDF, HDF, GeoTiff, ...
  - many metadata schemas
    - Biological Data Profile, Darwin Core, Dublin Core, **Ecological Metadata Language (EML)**, Open GIS schemas, ISO Schemas, ...
- Accepting this heterogeneity is critical

Metadata


knb
ABOUT
DATA
SHARE
TOOLS
Search for data
SIGN IN
‹ Back to search
Thissell A and Balch J. Testing Amazon Transitional Forest Leaf Flammability: Combustion Experiment Temperature Readings (2013) (doi:10.5063/F1XW4GQH)
Files
Package urn:uuid:19d860f4-7291-4cd0-a55c-5aba5f511662 .zip 14.79 MB 4 downloads
Metadata doi:10.5063/F1XW4GQH .xml (EML) 11.28 KB 79 views
Data jkbalch.14.1 text/csv 14.78 MB 6 downloads Details
Download all
General
Title Testing Amazon Transitional Forest Leaf Flammability: Combustion Experiment Temperature Readings (2013)
Identifier jkbalch.15.2
Abstract Intensifying land use practices and global climate change are leading to increased forest fragmentation and more severe droughts in the southeastern transitional forests of the Brazilian Amazon. The combination of land use and droughts has increased fire frequency, intensity, and extent in the region. Various plant traits can either limit or promote flammability, as demonstrated in many seasonally dry environments (e.g. Mediterranean shrublands), which are home to species that are adapted to frequent fire. However, little research has been done to associate flammability with the leaf traits of humid tropical forest species that are experiencing fire with greater frequency and severity than historically recorded. In this study, the leaf traits and consequent burning characteristics of seventeen species abundant in a transitional forest plot in Mato Grosso, Brazil were analyzed through controlled combustion experiments and trait measurements. Principal components analysis and linear regression were employed to compare leaf traits and their ability to predict flammability. The results reported herein show that these species demonstrate variability in flammability. Individual leaf surface area and leaf volume were found to be the most significant predictors of flammability among these species. The most flammable species were those with thin, lightweight leaves that become arranged in loosely packed, highly aerated fuelbeds and would be able to drive rapidly moving, high intensity fire across the forest floor, consuming a larger amount of fuel. The less flammable species were characterized by thick, heavy leaves that create dense fuelbeds, potentially inhibiting fire spread and the overall consumption of available fuels. Analyzing the dynamics between leaf traits and plant flammability with forest composition, successional trajectories and future fire behavior could lead to an examination of overall landscape effects in which certain species with fire-adaptive strategies could outcompete more fire-sensitive species, therefore influencing future fire regimes. Furthermore, given a scenario of increasing fire, this knowledge can lead to more effective conservation and land use policies in tropical regions.

Owner and Contact Metadata

People and Associated Parties

Data Set Creators	Individual	**Amy Thissell**
	Organization	**The Pennsylvania State University**
	Position	Primary Data Set Owner
	Individual	**Dr. Jennifer Balch**
	Organization	**University of Colorado**
	Position	Co-Owner
	Address	Guggenheim Building, 260 UBC, Room 110, University of Colorado, Boulder, Colorado 80309 USA

Data file metadata

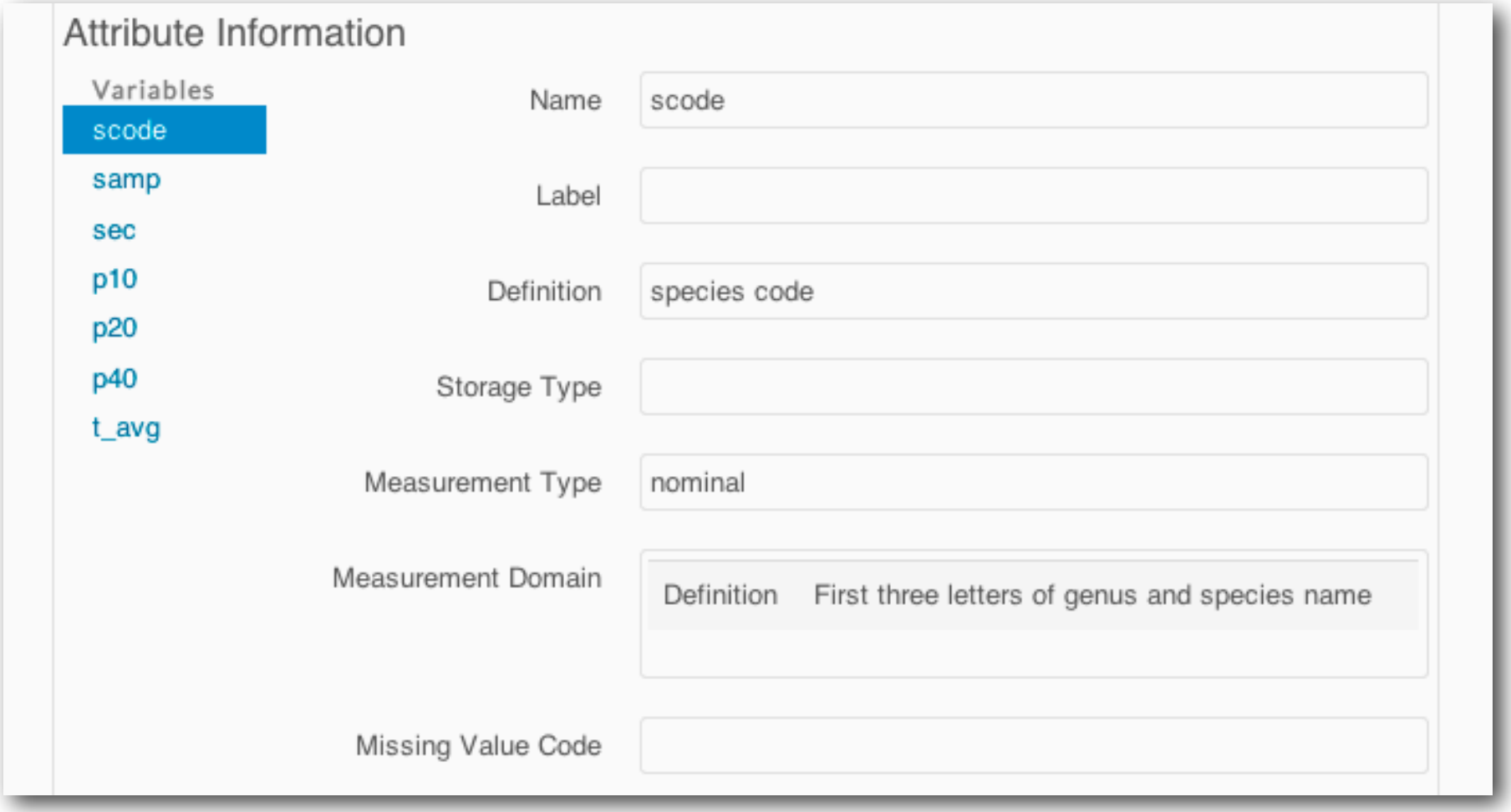
Data Table, Image, and Other Data Details

Technical Metadata	Ecological Metadata Language (EML) File	
Data Table	Entity Name	**alltemps_Thissell_2013.csv**
	Object Name	alltemps_Thissell_2013.csv
	Online Distribution Info	jkbalch.14.1
	Size	15500059 byte
	Text Format	

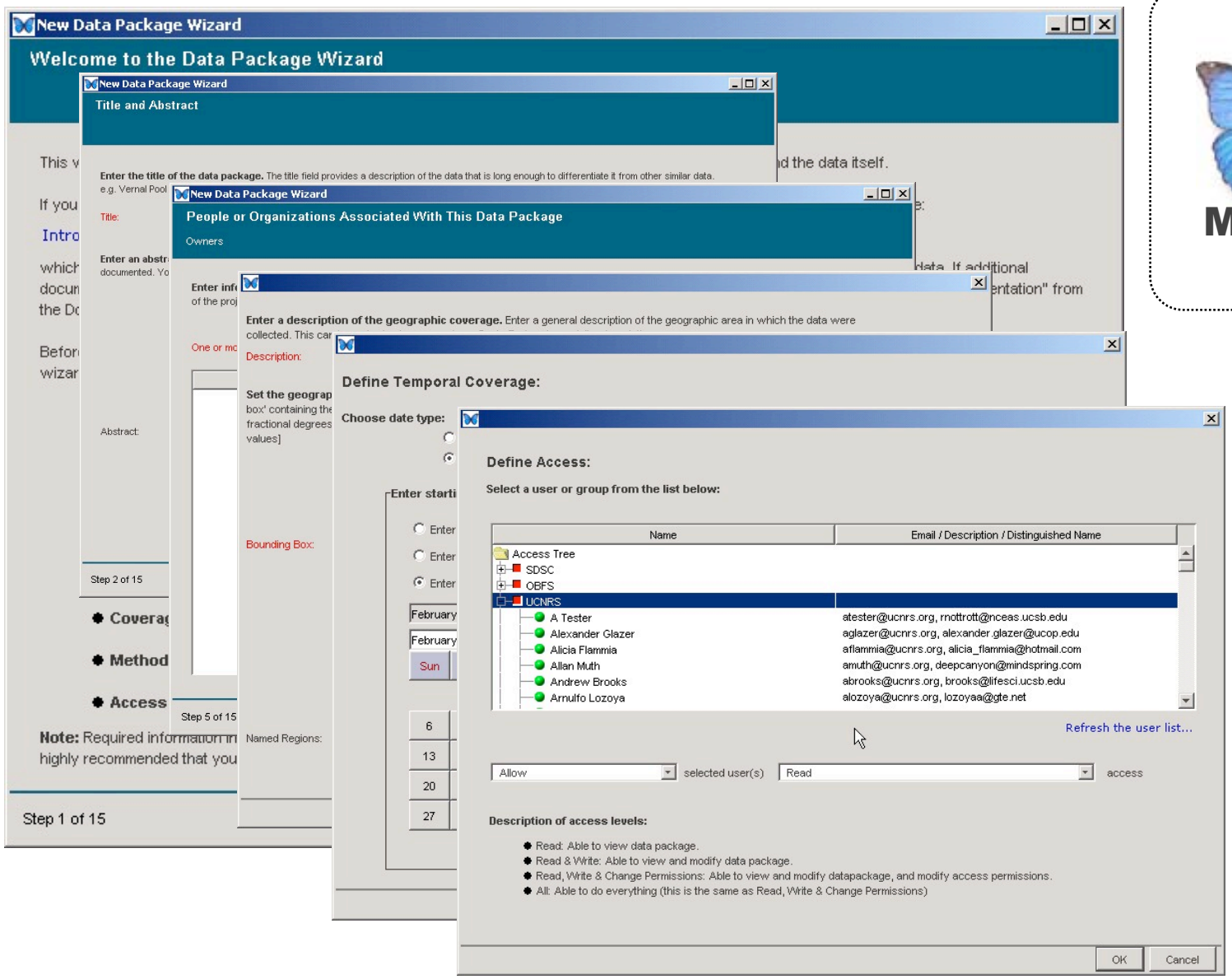
Number of Header Lines	1
Record Delimiter	#x0A
Attribute Orientation	column
Simple Text	
Field Delimeter	,

Number Of Records	214569

Column metadata



Wizards to create metadata



Morpho

Morpho highlights

- Create metadata in EML format
- Manage data in EML packages
- Save, publish, and share data

- Search for data
- Multi-language
  - English, Spanish, Chinese, French, **Portuguese**, Japanese
- Export data and metadata
- Cross-platform, and open source

Data Citation

Data Set Citation:

When using this data, please cite the data package:

Kline T.
SEA: Confirming Food Web Dependencies with Stable Isotope Tracers: Food Webs of Fishes-Prince William Sound, Alaska (1994-1998)
couture.9.15 (http://evos.nceas.ucsb.edu/evos/metacat/couture.9.15/default)

- NCEAS can issue DOI identifiers for publicly archived data sets:

  doi:10.xxxx/AA/gulfwatch.9.15

- Always resolve to the data set
- Used in journals to cite data usage

CREATE NETWORKS

Global Metacat deployments

LTER Data Catalog

The US
Long Term Ecological Research
Network

LTER Home
Login
Search
Browse

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

Data Set Citation

Drucker D of BR-LTER.**Abundância e Distribuição de Ervas Terrestres em Parcelas Ripárias na Reserva Ducke: Variação Lateral**.
drucker.3.1 (http://aranha.inpa.gov.br/knb/metacat/drucker.3.1/ppbio).

Metadata download: **Ecological Metadata Language (EML) File**

Data Set Owner(s):

Individual: **MSc. Debora Drucker**
Organization: BR-LTER
Email Address: deboradrucker@gmail.com

Abstract:

Os dados aqui disponibilizados são produto do trabalho realizado por Debora Drucker durante seu mestrado. O objetivo central foi investigar a abundância e distribuição espacial de ervas terrestres (apenas as espécies que germinam e passam todo o seu ciclo de vida no solo) em 20 Parcelas ripárias paralelas aos igarapés na Reserva Florestal Adolpho Ducke.

Keywords:

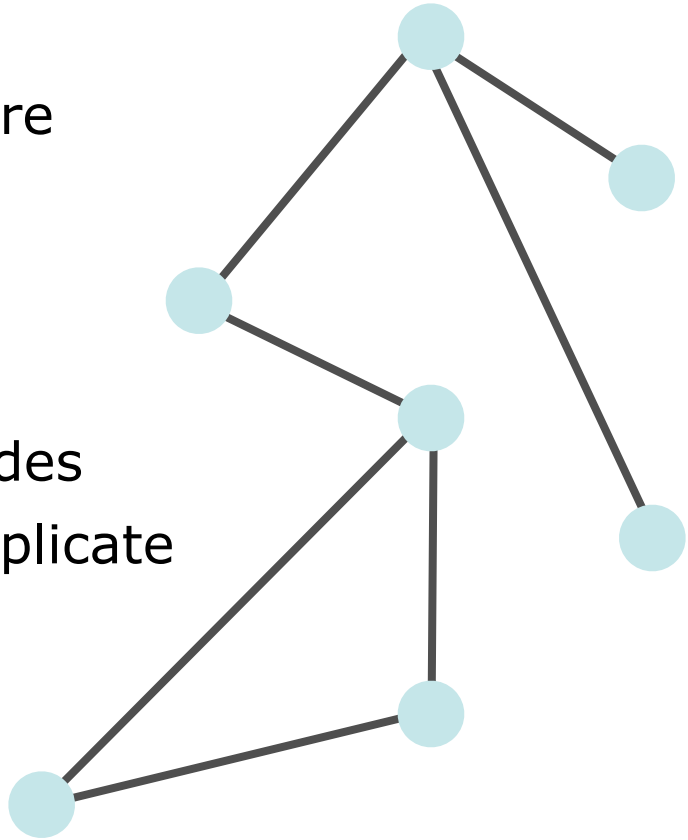
- Ervas
- Parcelas Ripárias
- Reserva Ducke
- Floresta de Terra Firme

License and Usage Rights:

A Federation of repositories

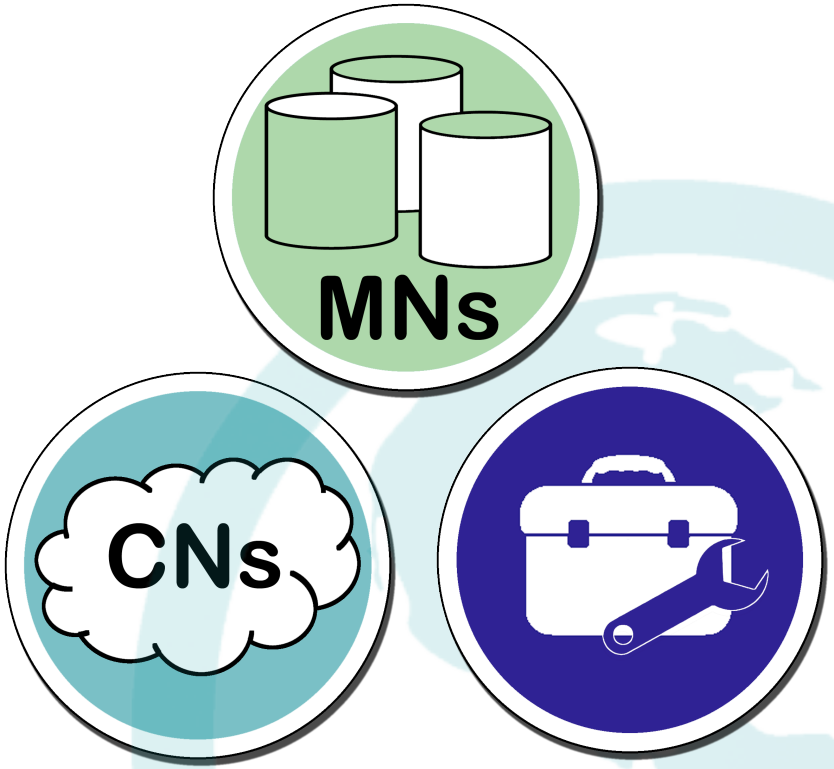
- **Diverse Federation == Resilience**
  - Failover for temporary outages
  - Insurance against institutional failure
  - Avoid data loss in the long-term

- **Diverse Federation == Scalability**
  - Storage increases with Member Nodes
  - Incremental costs to each MN to replicate
  - Distributes sustainability costs



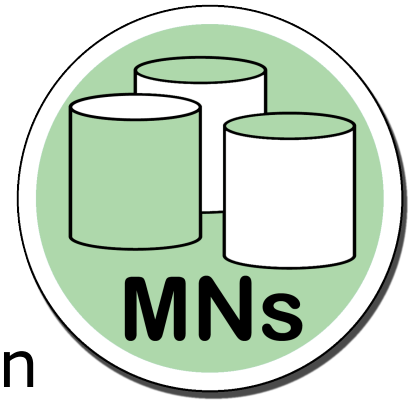
Creating Interoperability

- **Member Nodes (MNs)**
  - Heart of the federation
  - Harness the power of local curation
- **Coordinating Nodes (CNs)**
  - Services to link Member Nodes
- **Investigator Toolkit (ITK)**
  - Tools for the whole data lifecycle



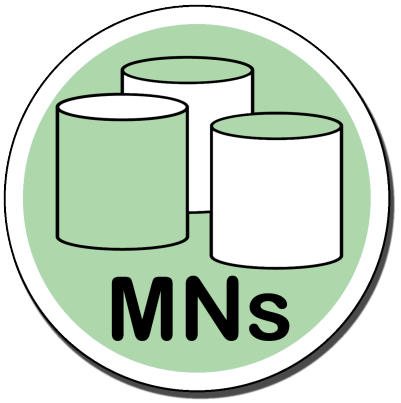
Interoperability

Member Nodes



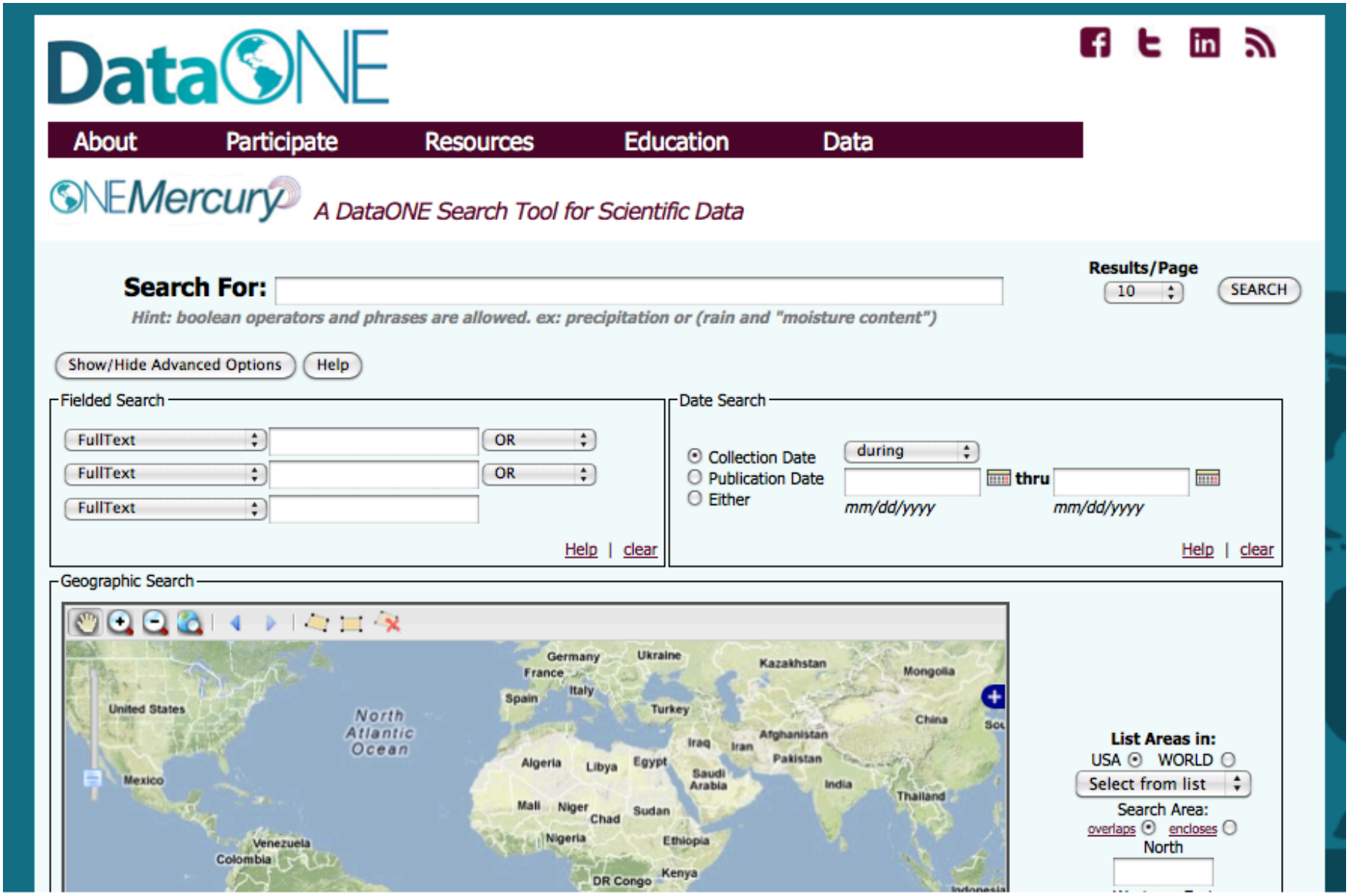
- **Authoritative** members of the Federation
- **Curate** data holdings
  - *Provide unique identifiers for each object*
  - *Ensure availability, quality, and reliability*
- **Replicate** holdings for other MNs
- Provide access and **access control**
- **Log** and report accesses to objects
- Engage with DataONE community
- Deploy DataONE-compatible software systems

Member Nodes



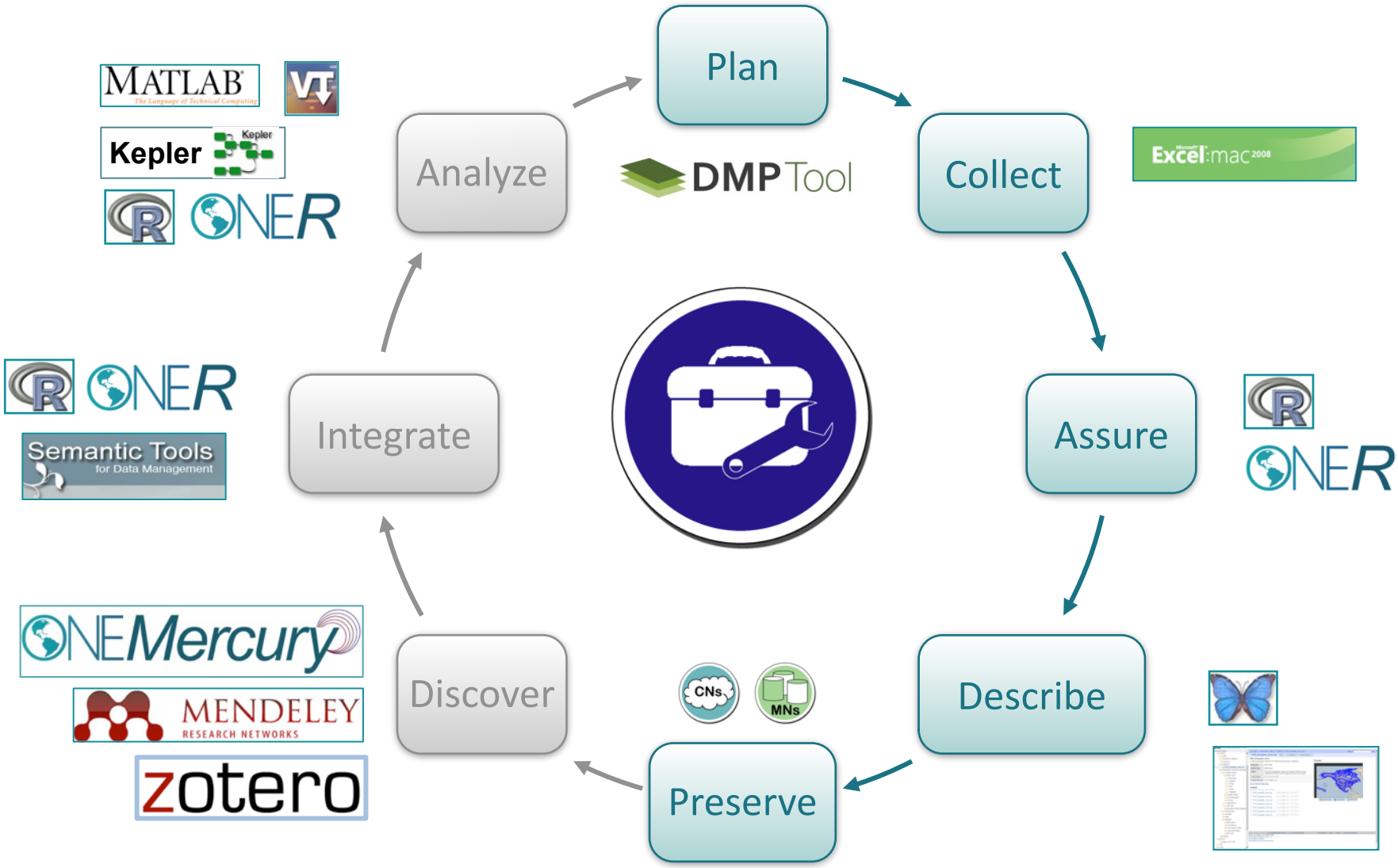
Avian Knowledge Network

...and many more!


DataONE
About
Participate
Resources
Education
Data
ONEMercury
A DataONE Search Tool for Scientific Data
Search For:
Hint: boolean operators and phrases are allowed. ex: precipitation or (rain and "moisture content")
Results/Page
10
SEARCH
Show/Hide Advanced Options
Help
Fielded Search
FullText
OR
FullText
OR
FullText
Help | clear
Date Search
Collection Date
Publication Date
Either
during
thru
mm/dd/yyyy
mm/dd/yyyy
Help | clear
Geographic Search
North Atlantic Ocean
United States
Mexico
Venezuela
Colombia
Germany
France
Spain
Italy
Ukraine
Turkey
Kazakhstan
Mongolia
China
Iraq
Iran
Afghanistan
Pakistan
India
Thailand
Algeria
Libya
Egypt
Saudi Arabia
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
DR Congo
Kenya
List Areas in:
USA
WORLD
Select from list
Search Area:
overlaps
encloses
North

CREATE INTEROPERABLE SOFTWARE

Software Interoperability



KNB System Components

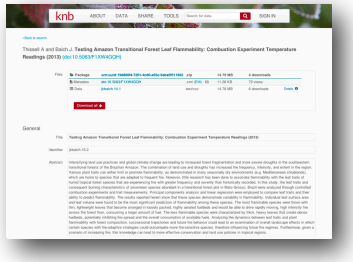
Document → **Share** → **Analyze** → **Communicate**

Morpho

Meta Cat

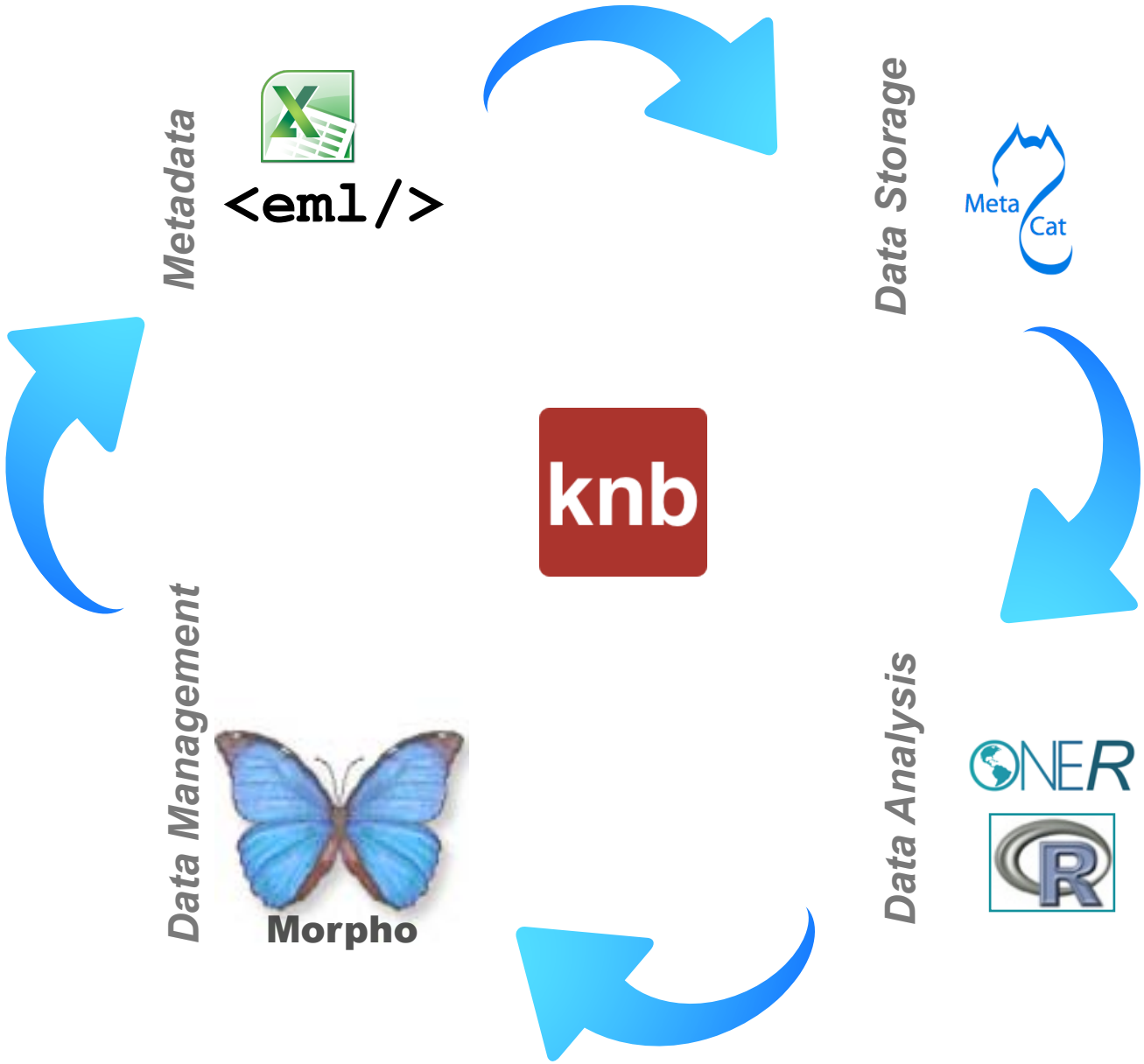
ONE*R*

W E B



Metadata → **Data** → **Workflows** → **Results**

Data life-cycle



How do we harness the long tail?

- Efficient data federation
  - Focus on individual contributors
- Late binding in informatics systems
  - Loose coupling
  - Schema-less storage
- Central search for discovery
- Interoperable software

Data Registration Activity

- https://identity.nceas.ucsb.edu/

- Register sample dataset
  - https://dev.nceas.ucsb.edu/#share

Questions?

- Contact:
  - Ben Leinfelder <leinfelder@nceas.ucsb.edu>
  - Matt Jones <jones@nceas.ucsb.edu>

- Links
  - http://www.nceas.ucsb.edu/ecoinfo/
  - https://knb.ecoinformatics.org/
  - http://www.dataone.org