Data sharing for scientific synthesis in ecology: challenges and opportunities

Eric Garnier (CNRS)
Director of CESAB

CESAB
CENTRE DE SYNTHÈSE ET D'ANALYSE SUR LA BIODIVERSITÉ

FRB
FONDATION POUR LA RECHERCHE SUR LA BIODIVERSITÉ

A centre created and developed by the FRB

cnrs
dépasser les frontières

Outline of the talk

- **What is CESAB and what is scientific synthesis?**
- **Challenges and solutions in the sharing of data**
  - **sociological and cultural**
  - **“technical”**
  - **semantical**

What is CESAB?

- The “Centre for the Synthesis and Analysis of Biodiversity”
- One of the five flagship programmes of the French Foundation for Research on Biodiversity (FRB)…
- … and a synthesis and analysis centre (created in 2010)
- **Objective**: promote high level research on a wide range of topics related to biodiversity, without collection of new primary data

What is CESAB?

- A centre to provide space and “timeless time” to advance knowledge on all fields pertaining to biodiversity

- Located in Aix-en-Provence (~ 30 km North of Marseille)

- Selection of 3 to 4 working groups/year based on a call for proposals

FORESTERRA

Enhancing FOrest RESearch in the MediTERRAnean through improved coordination and integration

The 10 current CESAB working groups

BETSI

TRAITS FONCTIONNELS, BIOLOGIQUES ET ÉCOLOGIQUES, D'INVERTÉBRÉS DU SOL. RÉPONSE DES ORGANISMES DU SOL AUX FACTEURS ENVIRONNEMENTAUX ET DÉVELOPPEMENT DE BIO-INDICATEURS

Porteur du projet : Mickaël Hedde, INRA Versailles- Grignon, France

NETSEED

AGROBIODIVERSITÉ ET RÉSEAUX SOCIAUX : COMMENT LES SYSTÈMES D'ÉCHANGES DE SEMENCES AGISSENT SUR LA DIVERSITÉ DES PLANTES DOMESTIQUÉES

Porteur de projet : McKey Doyle, CEFE – CNRS- INEE Université Montpellier II, France

BIODIS

DÉCRYPTER LES LIENS ENTRE BIODIVERSITÉ ET MALADIES INFECTIEUSES ÉMERGENTES

Porteur du projet : Jean-François Guégan, IRD - UMR MIVEGEC, Montpellier, France ; Co-porteurs : Benjamin Roche (IRD - UMMISCO, Paris) et James N. Mills (Emory University, Atlanta, USA)
contact: jean-francois.guegan@ird.fr

DIVGRASS

DIVERSITÉ VÉGÉTALE ET FONCTIONNEMENT DES PRAIRIES PERMANENTES

Porteur du projet : Philippe Choler, Laboratoire d'Ecologie Alpine – CNRS-INEE, Université J. Fourier, Grenoble, France

ISLANDS

ASSEMBLAGE DES COMMUNAUTÉS ÉCOLOGIQUES SUR LES ÎLES LOINTAINES: VERS UN NOUVEAU MODÈLE DE BIOGÉOGRAPHIE INSULAIRE?

Principal Investigator: Christophe Thébaud, Université Paul Sabatier, Toulouse , France

LOLA-BMS

COMMENT LES PROCESSUS LOCAUX EXPLIQUENT-ILS LA RÉPONSE DES PAPILLONS AUX CHANGEMENTS PLANETAIRES : UNE ANALYSE INTÉGRÉE À PARTIR DES PROGRAMMES DE SUIVIS

Porteur de projet: Romain Julliard, MNHN, France, **Co-Porteur:** Guy Pe'er, UFZ, Allemagne; contact: lola-bms@cesab.org

IRBAS

ANALYSE ET SYNTHÈSE DE LA BIODIVERSITÉ DES RIVIÈRES INTERMITTENTES

Porteur du projet : Thibault Datry, IRSTEA, Lyon
contact: irbas@cesab.org

GASPAR

RELATIONS « DIVERSITÉ- ABONDANCE », UNE CLÉ POUR COMPRENDRE LES CONSÉQUENCES DES CHANGEMENTS GLOBAUX SUR LES ÉCOSYSTÈMES : LES POISSONS CORALLIENS COMME MODÈLE

Porteur de projet : Michel Kulbicki, Laboratoire Arago, IRD, Baynuls/mer, France

AFROBIODRIVERS

DYNAMIQUE DE LA BIODIVERSITÉ AFRICAINE: INTER-ACTIONS ENTRE PROCESSUS ÉCOLOGIQUES ET ACTIONS DE CONSERVATION

Porteur de projet: Hervé Fritz (Hervé.Fritz@univ-lyon1.fr), Laboratoire Biométrie et Biologie évolutive, CNRS, UCB Lyon 1, Villeurbanne

PELAGIC

ETABLIR DES AIRES PERTINENTES D'UN POINT DE VUE ÉCOLOGIQUE ET IMPORTANTES À L'ÉCHELLE PLANÉTAIRE POUR LA CONSERVATION DES MAMMIFÈRES MARINS : UNE SYNTHÈSE DES MEILLEURES CONNAISSANCES DISPONIBLES POUR INFORMER LES POLITIQUES DE GESTION

Porteur de projet: David Kaplan, IRD, France
contact: pelagic@cesab.org

MORE INFORMATION ON

www.cesab.org

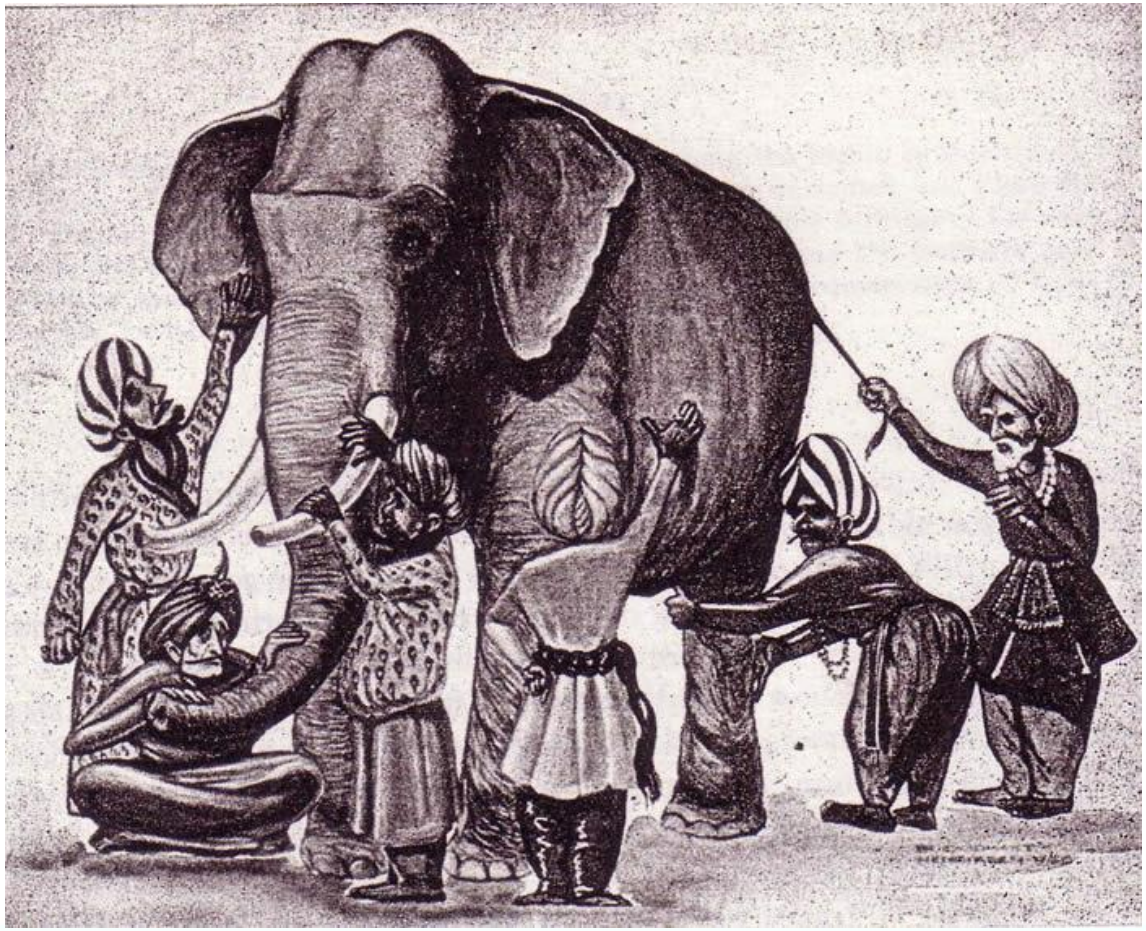
What is scientific synthesis?

- Scientific synthesis refers to the integration of diverse research in order to increase the generality and applicability of the results of that scientific research

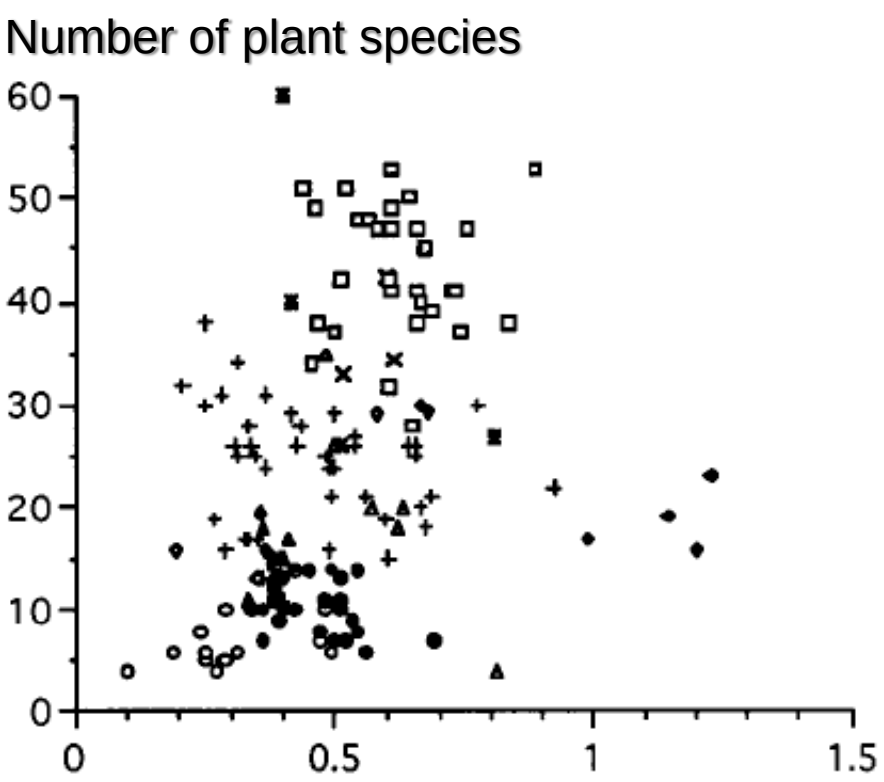
- Synthesis occurs both within and across disciplines and professional sectors and is therefore not captured entirely by the term *interdisciplinary research*

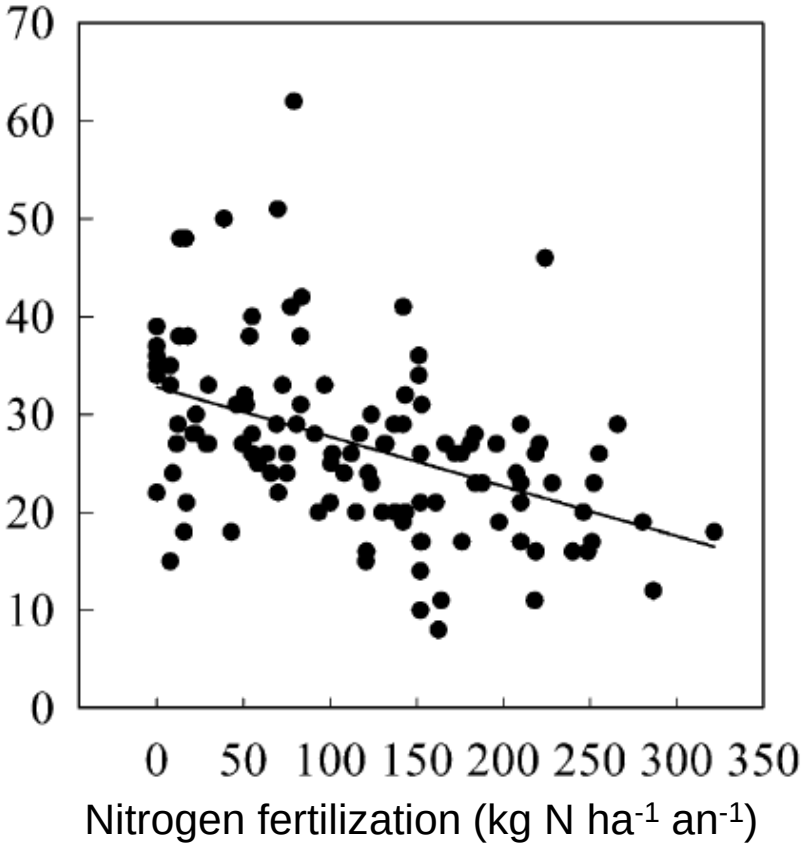
Hampton & Parker (2011)
BioScience 61: 900

Why care about scientific synthesis?



An example: how does fertilization affect plant diversity in grasslands?





Janssens *et al.* (1998) *Plant Soil* 202: 69

Klimek *et al.* (2007) *Biol Cons* 134: 559

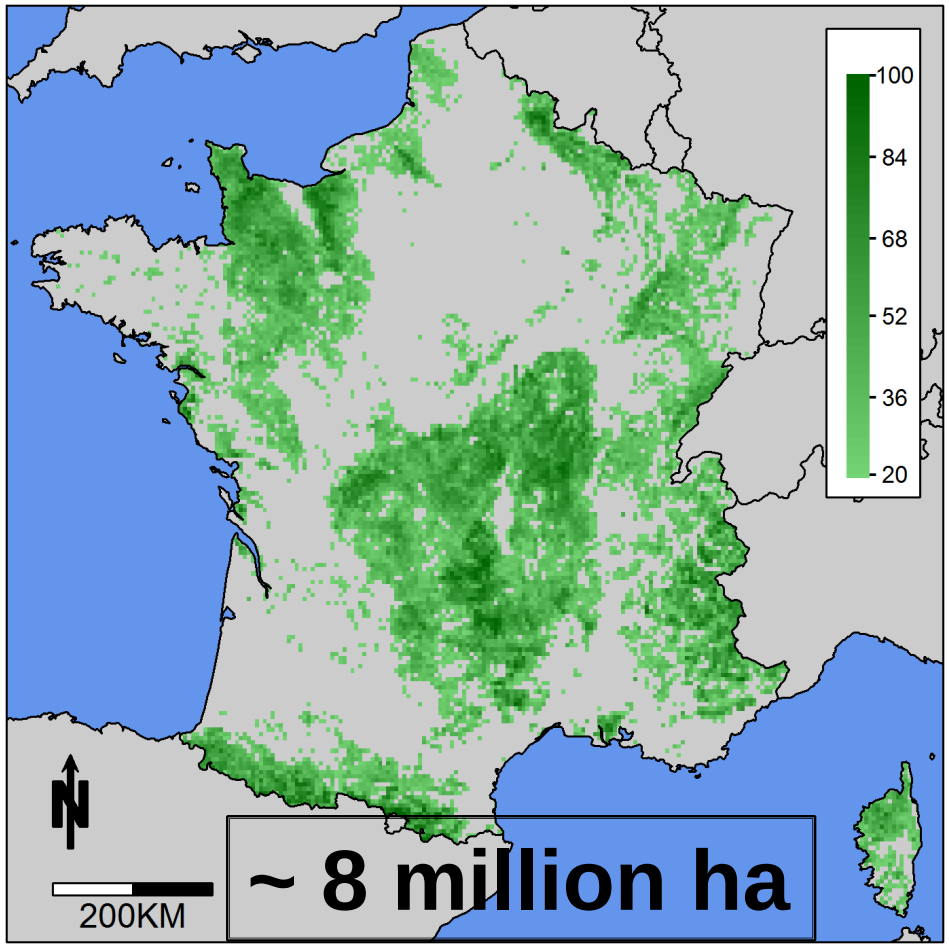
The basic pieces required to understand this relationship

- **Definitions of concepts:**
  - biodiversity
  - fertility
- **Choice of descriptive variables:**
  - number of species (biodiversity index)
  - components of nitrogen availability (indicator of fertility)
- **Methods**
- **Interactions with other factors:**
  - biogeographical context
  - soil-climate context
  - other resources (phosphorus, water…)
  - other management practices
  - …

SHARING DATA:

CHALLENGES AND SOLUTIONS

An example: the CESAB project DIVGRASS (DIVersity of GRASSlands)



What are the patterns of plant functional diversity in permanent grasslands along environmental gradients?

The data used in the context of DIVGRASS

Nature of data	Source	Access	Conditions
Floristic relevés	Members of WG	Free in the context of project	Free
	Botanical Conservatories	Agreement limited to the project	Free
	InfoSols – RMQS	Agreement limited to the project	Free
	SOPHY	(Very) complex	Exchange
Plant traits	Members of WG	Free in the context of project	Free
	TRY data base	Proposal submitted to steering committee	Agreement with data custodians (TRY IP)
Taxonomy	TaxRef v4.0	Free (MNHN)	Free
Soil	Members of WG	Free in the context of project	Free
	InfoSols – RMQS	Agreement limited to the project	Free
	InfoSols – BDGFS	Agreement limited to the project	Free
Climate	Members of WG	Free in the context of project	Free
	Aurhély (Météo France)	Via one member of project	Free
Land use	Référentiel parcellaire graphique	To be paid for – Usable in context of the project	5500 € TTC
	Agreste (Ministry of Agriculture)	Free	Free

Challenges associated with scientific synthesis and data sharing

- **Technological: data**
- **Semantics: concepts**
- **Cultural and sociological: benefits of sharing and intellectual properties**

Reichman *et al.* (2011)
Science 331: 703

The data challenges

- **Data are dispersed:**
  - The vast majority of data in ecology is structured in small and dispersed sets of data, managed by « independent » researchers (less than 1% of the data are « freely » available)
  - The themes that have led to the collection of data might be substantially different

- **Data are heterogeneous**
  - Various sub-disciplines : *e.g.* organisms/communities/ecosystems; plants/animals/microbes
  - Lack of concertation among scientits to standardize approachs, protocols and data
  - Related fields (*e.g.* climatology, social sciences) have their own terminologies and experimental protocols
  - Syntactic heterogeneity

Possible solutions

Madin *et al.* (2008) *TREE* 23: 159

Semantics standards

- **Metadata:** who, what, when, where and how about every aspect of the data
  (*e.g.* Darwin Core, EML, ISO 19115 [INSPIRE])

- **Controlled vocabularies and thesaurus:**
  list of key terms and their definitions (in a domain of interest) and how these are organized and structured

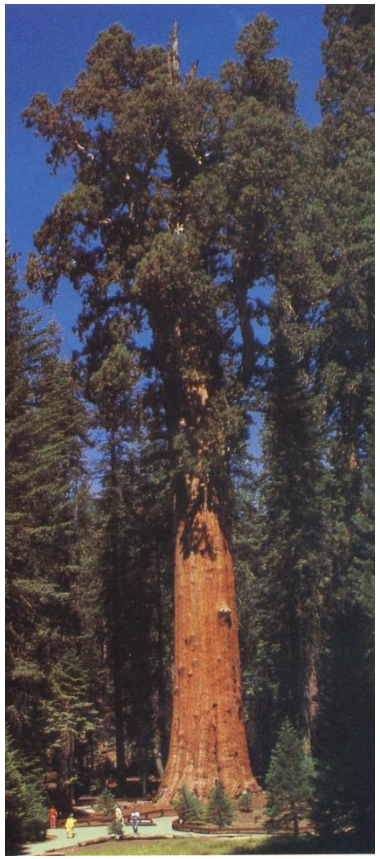
- **Ontologies:**
  a formal representation or classification of concepts and their relationships within a domain of interest

Example of a controlled vocabulary and a thesaurus for plant traits

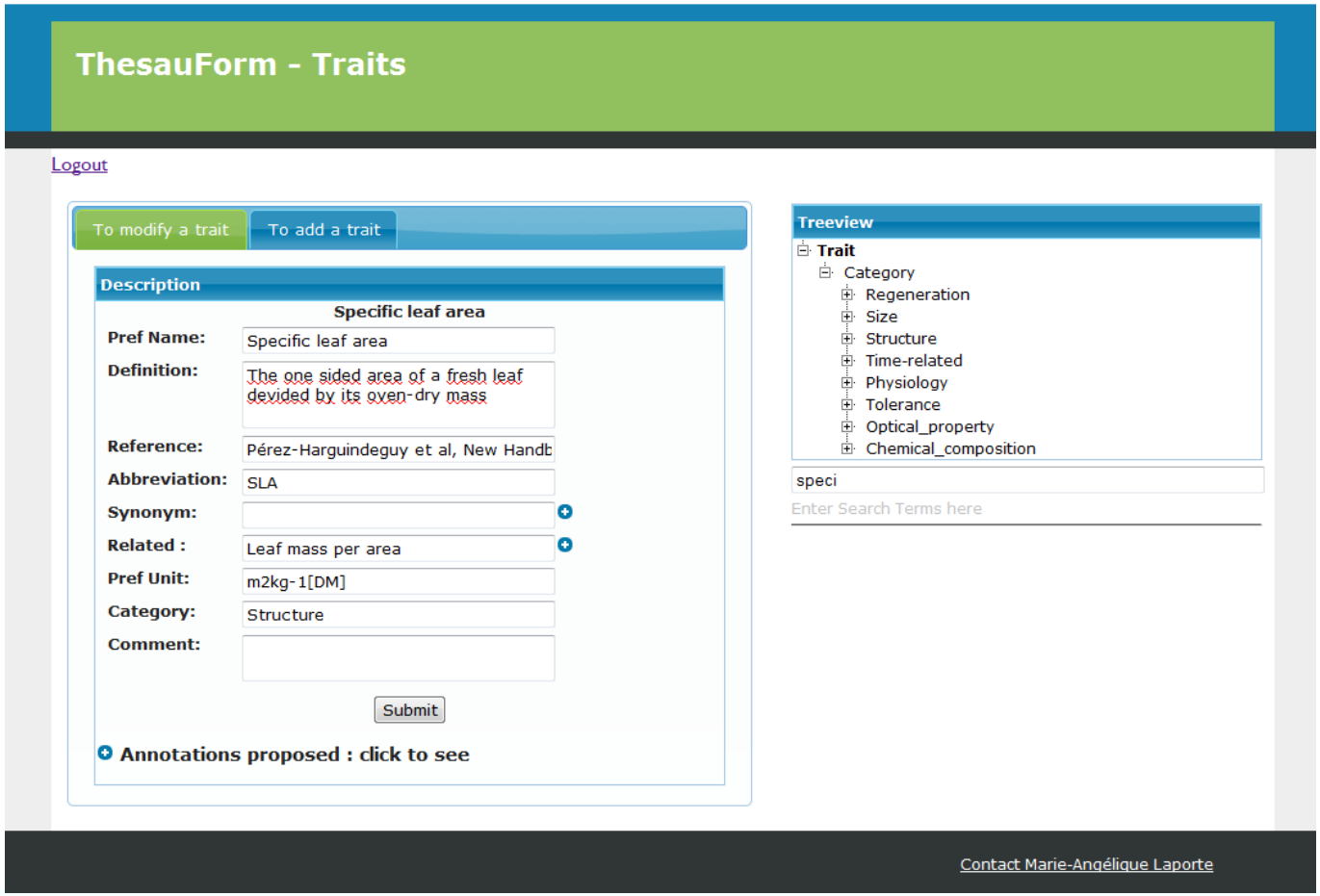
THE LEDA TRAITBASE

TRY

Plant height **Canopy height**

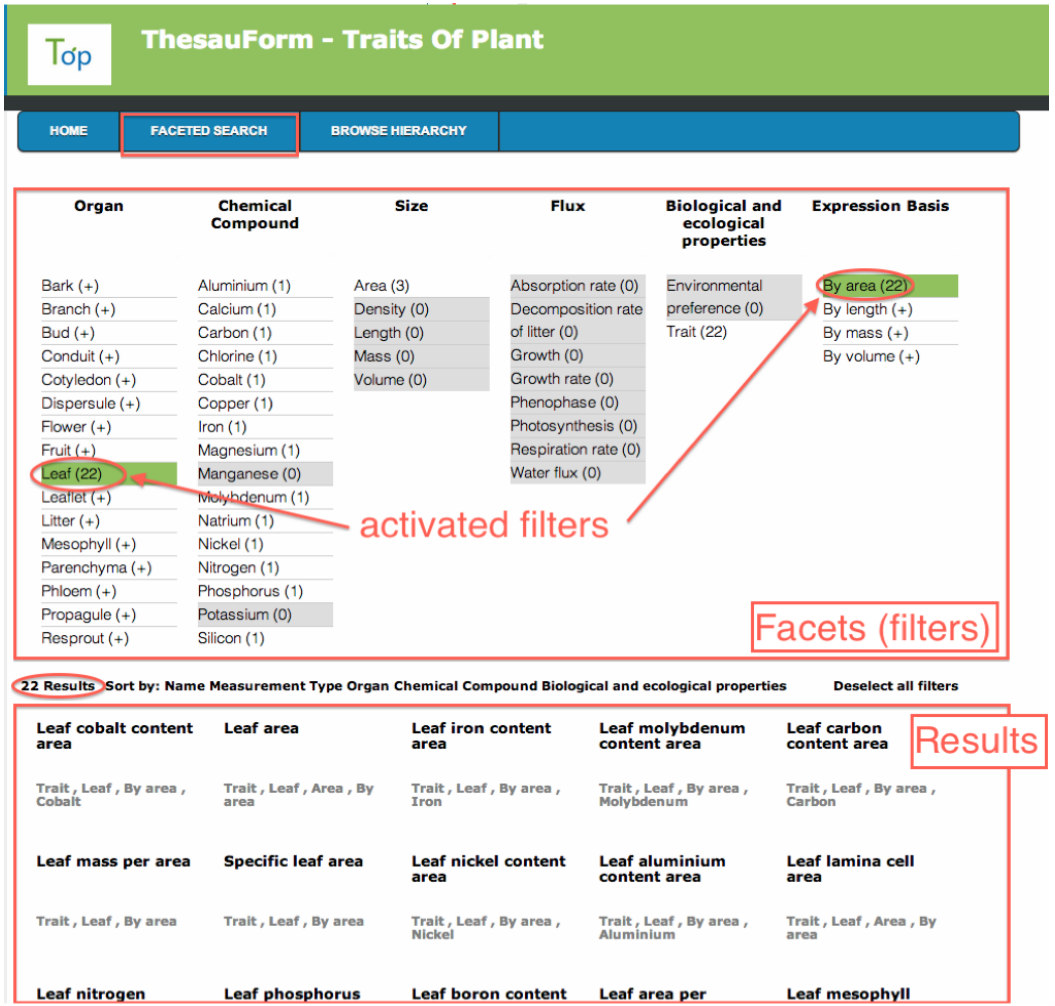


ThesauForm: a web tool for the collaborative construction of a thesaurus on plant traits



Laporte *et al.* (2012) *Ecol Inform* 11: 34

Visualizing the thesaurus (~ 1000 traits) : A facetted search tool



Laporte *et al.* (2013)
Proc S4Biodiv: ceur-ws.org/ Vol-979

Ontologies

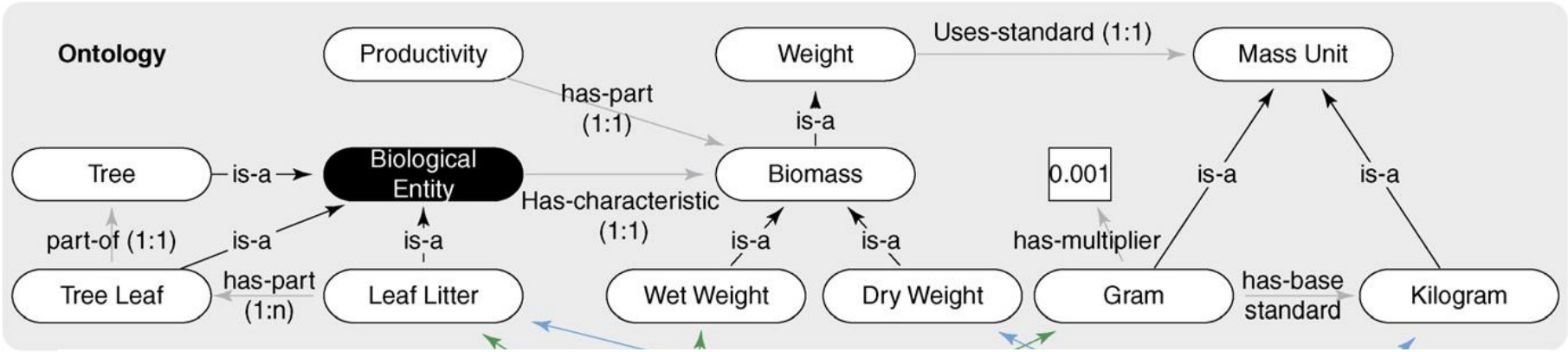
- In the information sciences, an ontology is a fixed universe of discourse in which:
  - each element or concept (*e.g.* field name or column in a database) is precisely defined
  - each possible relationship between data elements is parametized or constrained.(*e.g.* « is_a », « part_of », « has_member », « has_characteristic »…)

Schuurman & Leszczynski (2008) *BBI* 2: 187

- Explicit representation of a domain allowing ***a machine*** to execute automatically certain tasks implying some reasoning

- Must be ***shared*****,** and constructed in a way which makes it possible to improve it and add further concepts to it

A sample of an ontology in plant ecology



« **INTEROPERABILITY** » **OF DATABASES**
capacity of a system to operate with other systems without restrictions on access or implementation

The Holy Grail !

TRENDS in Ecology & Evolution

Madin *et al.* (2008)
TREE 23: 159

Cultural and sociological challenges (1)

- **Little incentive to share:**
  - data sharing is traditionally little developped in ecology as compared to other disciplines (*e.g.* astronomy, oceanography, genomics): analysis and publications of independent data sets
  - syntheses in ecology which have appeared in recent years and increased data flow lead to a change in this appraoch to data: advancement of science greatly benefit from data sharing (cf. NCEAS… and CESAB!)

- **Protection against « data predators »:**
  - retain data until they are not « correctly » valued, so that they are not used by others who could « steal » their originality and novelty
  - protection of intellectual property with penalties by peers or funding agencies still insufficient

Cultural and sociological challenges (2)

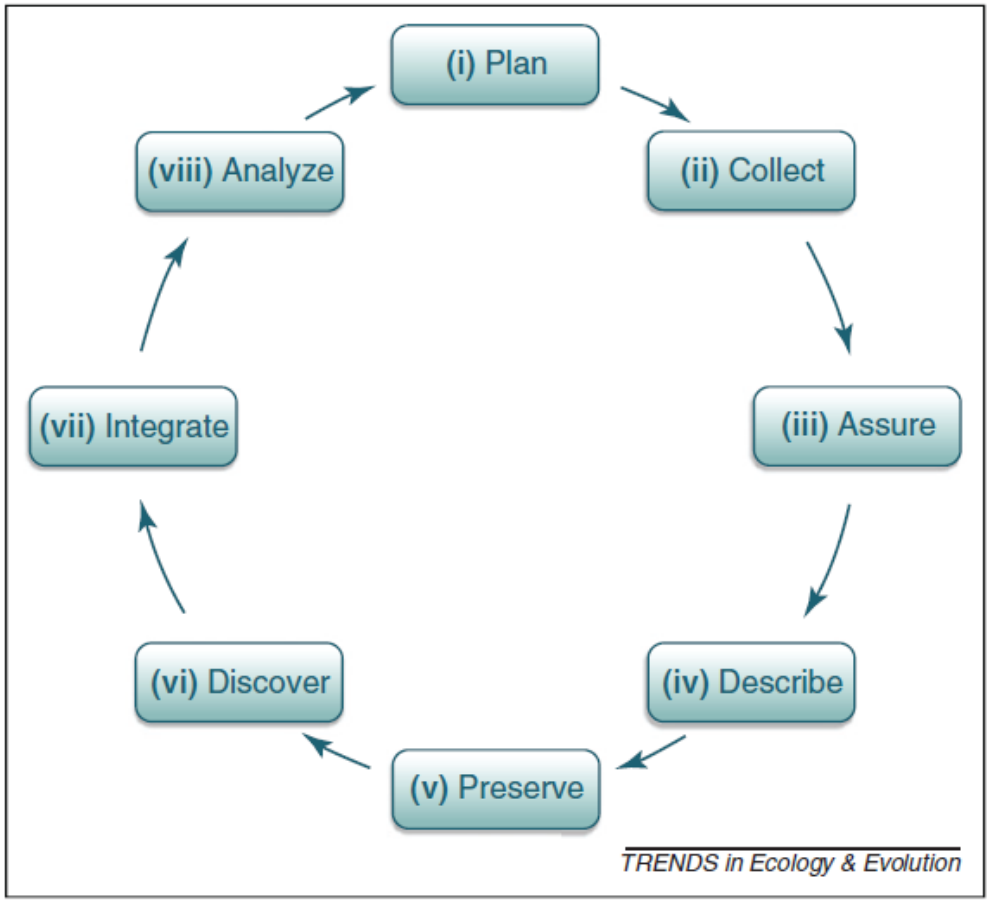
- **Reward:**
  - lack of reward for collecting data
  - make data sets publishable (« ecological archives » from ESA; journals PhytoKeys, ZooKeys, Nature), so that they are respected and valued as such, and reward those who make the effort (criteria for the selection of research projects)

- **Resources (human and financial) for the management of data:**
  - lack of funds
  - beyond individuals and specific projects: the scientific community and stakeholders should develop a perennial model for data management (cf. DataONE; GenBank) => needs funding!

CONCLUSIONS:

THE LIFE CYCLE AND MANAGEMENT PLAN OF DATA

Taking care of the data life cycle



Michener & Jones (2012)
TREE 27: 85

Implementing a data management plan

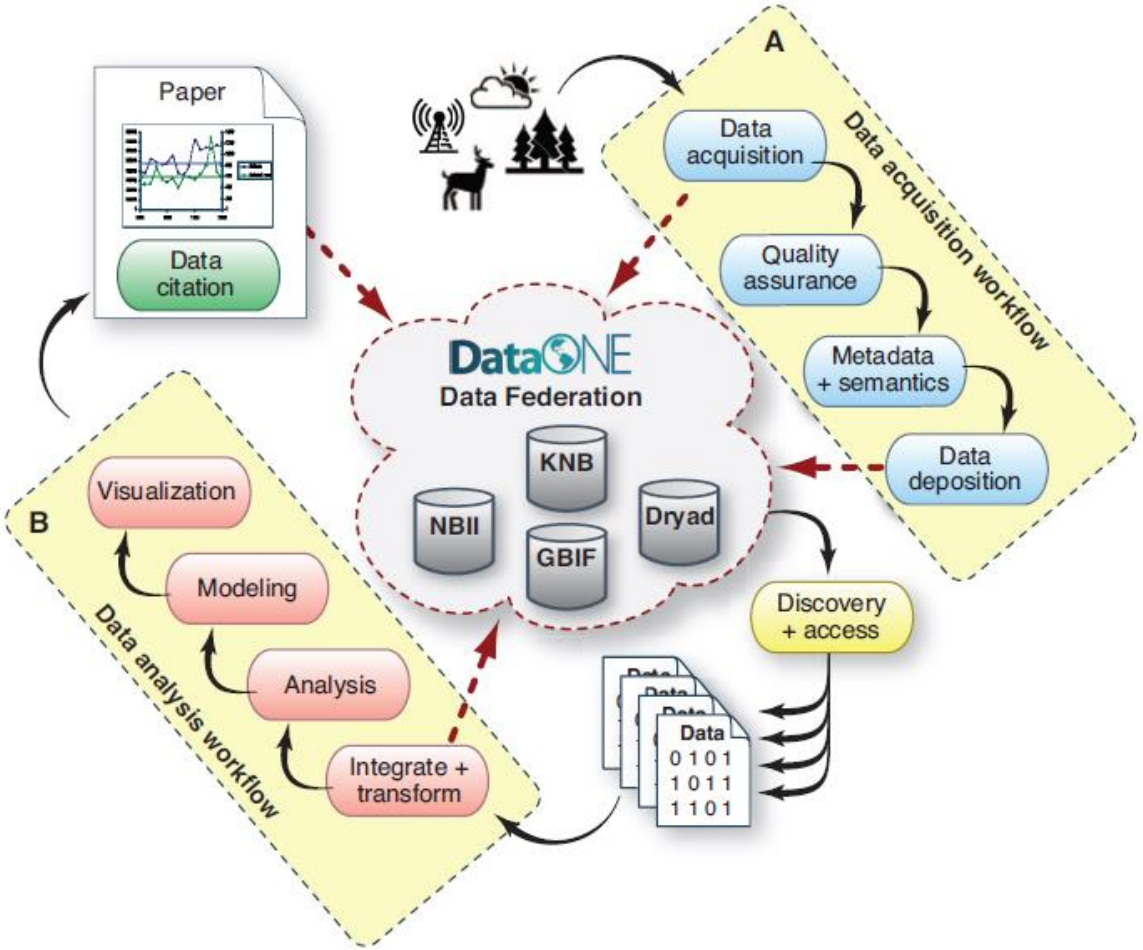
Component	Description and examples
Information about data and data format	Types of data that will be produced (e.g. experimental, observational, raw or derived, physical collections, models, images, etc.)
	When, where and how the data will be acquired (e.g. methods and instruments used)
	How the data will be processed (e.g. software, algorithms and workflows)
	File formats (e.g. csv, tab-delimited or naming conventions)
	QA/QC procedures used
	Other sources of data (e.g. origins, relationship to one's data and data integration plans)
	Approaches for managing data in the near-term (e.g. version control, backing up, security and protection, and responsible party)
Metadata content and format	Metadata that are needed
	How metadata will be created or captured (e.g. lab notebooks, auto-generated by instruments, or manually created)
	Format or standard that will be used for the metadata (e.g. EML or ISO 19115)
Policies for access, sharing and re-use	Requirements for sharing (e.g. by research sponsor or host institution)
	Details of data sharing (e.g. when and how one can gain access to the data)
	Ethical and privacy issues associated with data sharing (e.g. human subject confidentiality or endangered species locations)
	Intellectual property and copyright issues
	Intended future uses for data
	Recommendations for how the data can be cited (e.g. citation and DOI)
Long-term storage and data management	Identification of data that will be preserved
	Repository or data center where the data will be preserved
	Data transformations and formats needed (e.g. data center requirements and community standards)
	Identification of responsible parties
Budget	Anticipated costs (e.g. data preparation and documentation, hardware and software costs, personnel costs and archive costs)
	How costs will be paid (e.g. institutional support or budget line items)

Michener & Jones (2012) *TREE* 27: 85

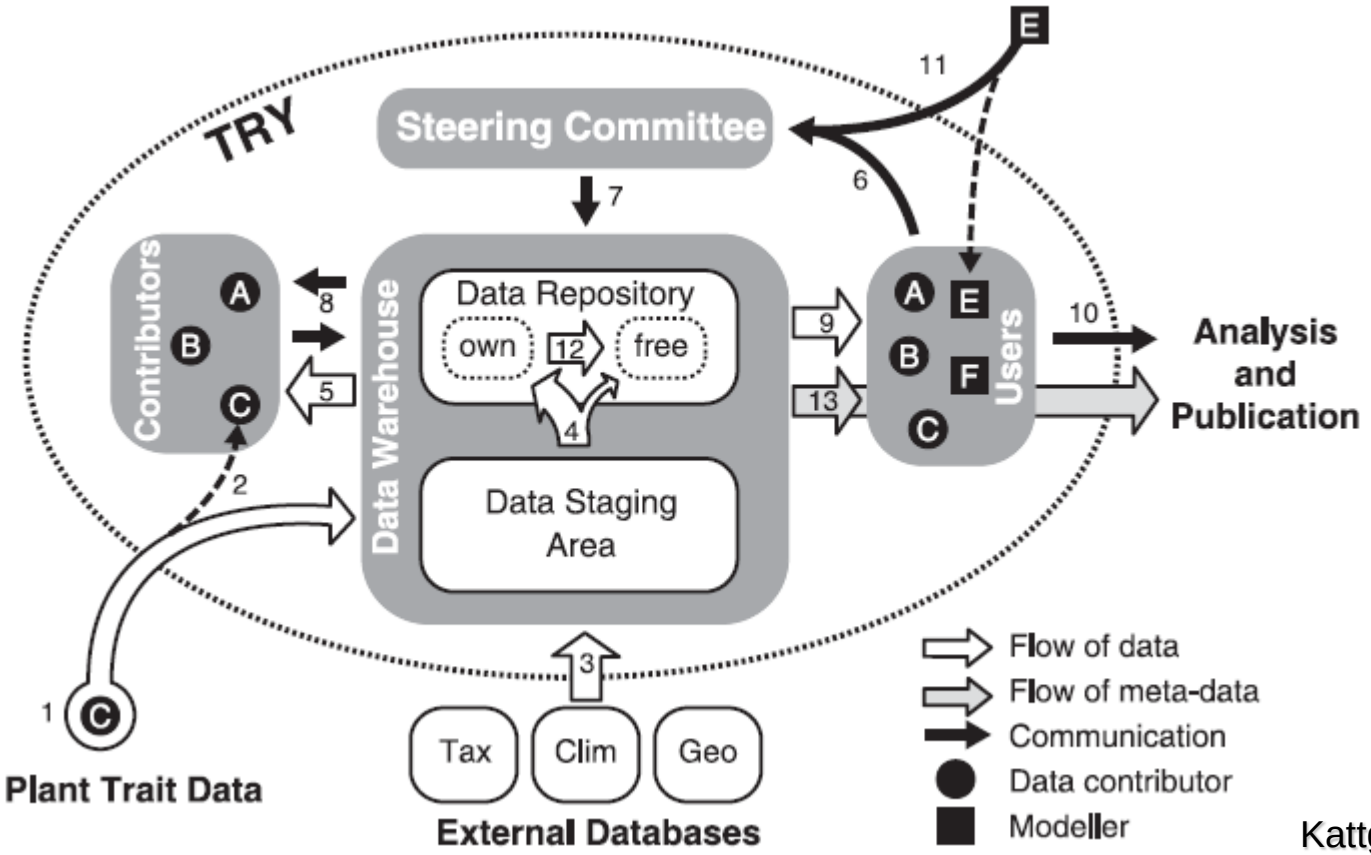
Thank you for your attention

Any question?

DataOne


Paper
Data citation
A
Data acquisition
Quality assurance
Metadata + semantics
Data deposition
Data acquisition workflow
DataONE
Data Federation
KNB
NBII
GBIF
Dryad
Discovery + access
Data
0 1 0 1
1 0 1 1
1 1 0 1
B
Visualization
Modeling
Analysis
Integrate + transform
Data analysis workflow

Data flow in TRY



Kattge et al. (2011)
GCB 17: 2905