Importance de la sémantique pour la synthèse des données en écologie

Eric Garnier

CENTRE D'ECOLOGIE FONCTIONNELLE & EVOLUTIVE

UMR 5175, Montpellier

Données et écologie

Data-intensive Science: A New Paradigm for Biodiversity Studies

STEVE KELLING, WESLEY M. HOCHACHKA, DANIEL FINK, MIREK RIEDEWALD, RICH CARUANA, GRANT BALLARD, GILES HOOKER

ZooKeys 150: 15–51 (2011)
doi: 10.3897/zookeys.150.1766
www.zookeys.org

REVIEW ARTICLE

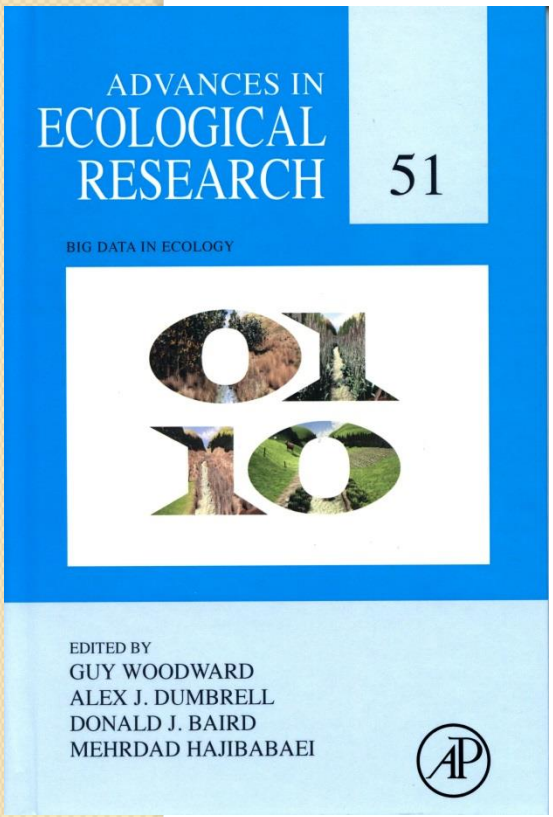
A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

GUEST EDITORIAL GUEST EDITORIAL GUEST EDITOR

Ecological data in the Information Age

Data issues in the life sciences

Anne E. Thessen, David J. Patterson



Biodiversity data should be published, cited, and peer reviewed

Mark J. Costello[1], William K. Michener[2], Mark Gahegan[3], Zhi-Qiang Zhang[4], and Philip E. Bourne[5]

CONCEPTS AND QUESTIONS

Big data and the future of ecology

Stephanie E Hampton[1*], Carly A Strasser[2], Joshua J Tewksbury[3], Wendy K Gram[4], Amber E Budden[5], Archer L Batcheller[6], Clifford S Duke[7], and John H Porter[8]

Journal of Vegetation Science **27** (2016) 865–867

COMMENTARY

Vegetation science in the age of big data

Scott L. Collins

EDITORIAL

Making Data Maximally Available

INTRODUCTION

Challenges and Opportunities

PERSPECTIVE

Challenges and Opportunities of Open Data in Ecology

O. J. Reichman,* Matthew B. Jones, Mark P. Schildhauer

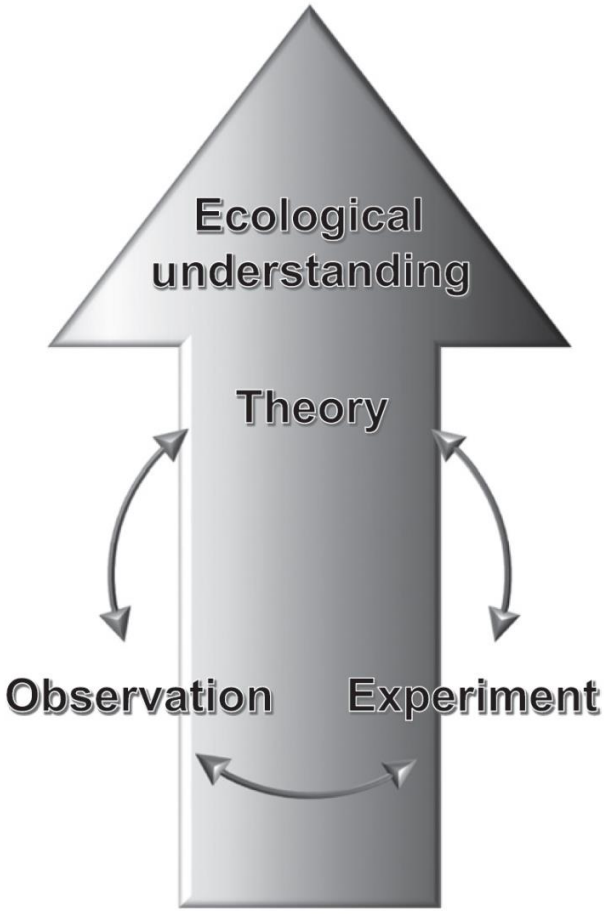
Donnée: ce qui est connu et admis, et qui sert de base, à un raisonnement, à un examen ou à une recherche

Organisation de l'exposé

- Pourquoi cet intérêt?
- Caractéristiques des données en écologie: les « données obscures »
- Exemples d'obstacles terminologiques à la synthèse des données

Pourquoi cet intérêt?

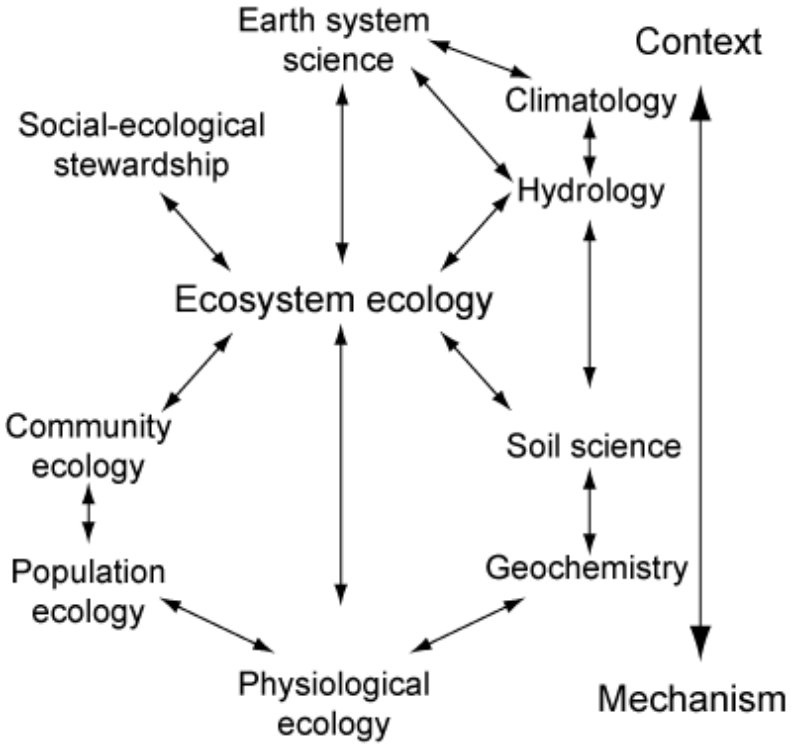
- Les données sont l'un des fondements des avancées scientifiques
- L'écologie est une discipline intégrative
- Explosion de données dans certains champs de l'écologie (« big data »)
- Certaines des questions posées à l'écologie ont une dimension continentale et/ou planétaire
- Développement des activités de synthèse scientifique



Eisenhauer *et al.* (2016) *JVS* 27 : 1061

Pourquoi cet intérêt?

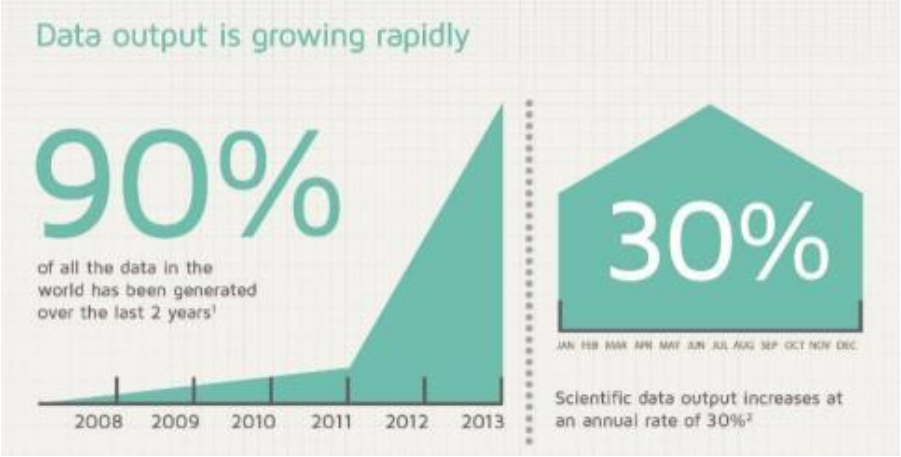
- Les données sont l'un des fondements des avancées scientifiques
- L'écologie est une discipline intégrative
- Explosion de données dans certains champs de l'écologie (« big data »)
- Certaines des questions posées à l'écologie ont une dimension continentale et/ou planétaire
- Développement des activités de synthèse scientifique

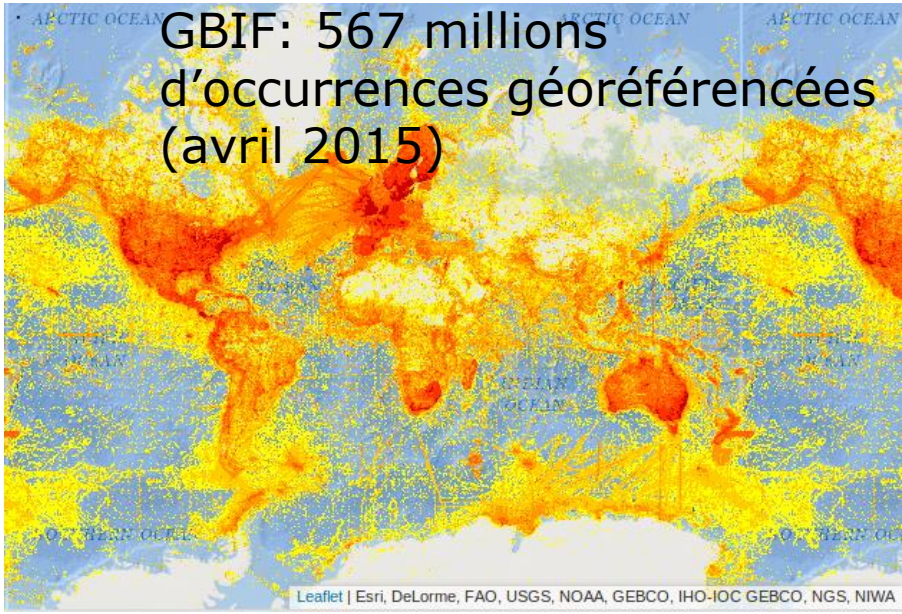


Chapin *et al.* (2011) *Principles of Terrestrial Ecosystem Ecology*

Pourquoi cet intérêt?

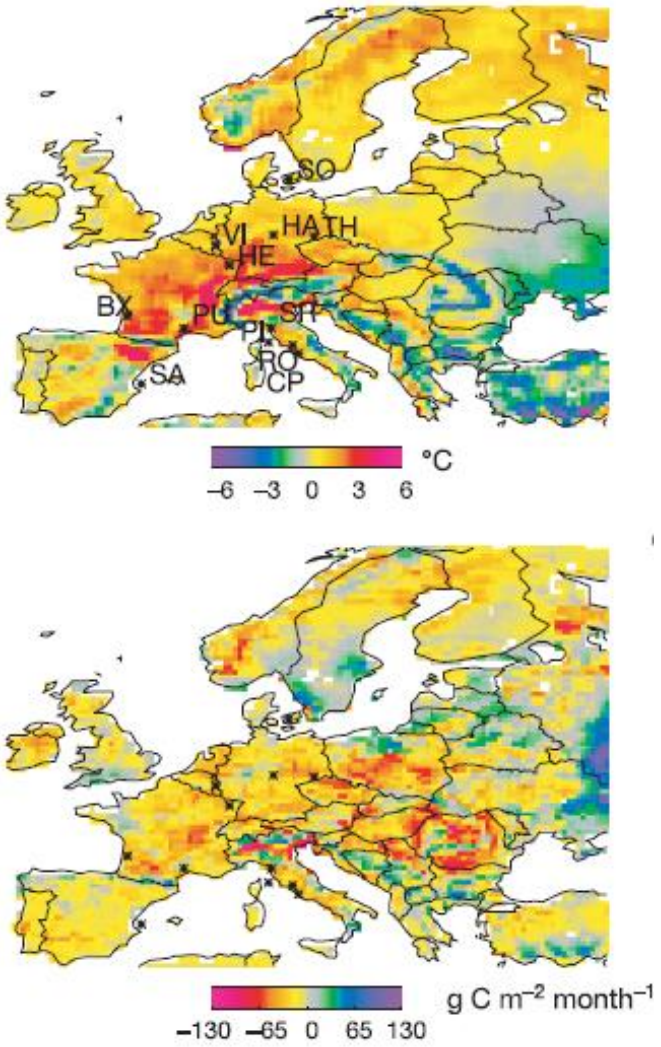
- Les données sont l'un des fondements des avancées scientifiques
- L'écologie est une discipline intégrative
- Explosion de données dans certains champs de l'écologie (« big data »)
- Certaines des questions posées à l'écologie ont une dimension continentale et/ou planétaire
- Développement des activités de synthèse scientifique





Pourquoi cet intérêt?

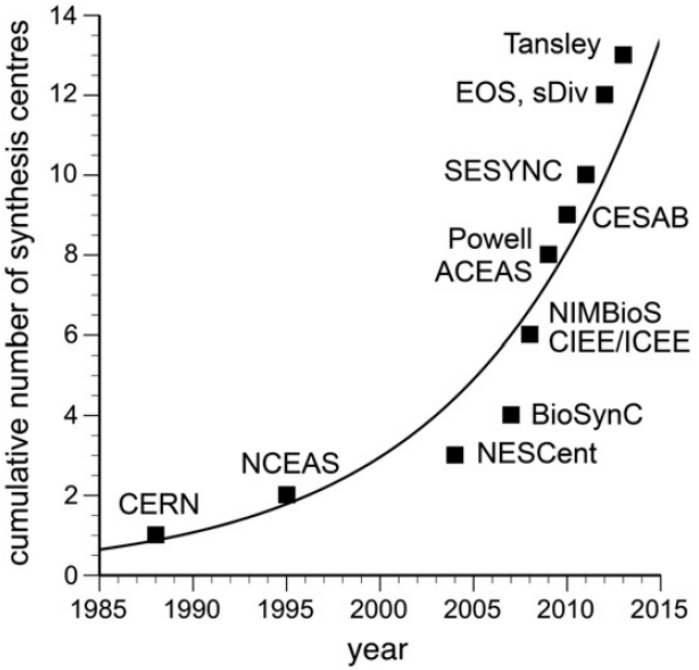
- Les données sont l'un des fondements des avancées scientifiques
- L'écologie est une discipline intégrative
- Explosion de données dans certains champs de l'écologie (« big data »)
- Certaines des questions posées à l'écologie ont une dimension continentale et/ou planétaire
- Développement des activités de synthèse scientifique



Ciais *et al.* (2005) *Nature* **437** : 529

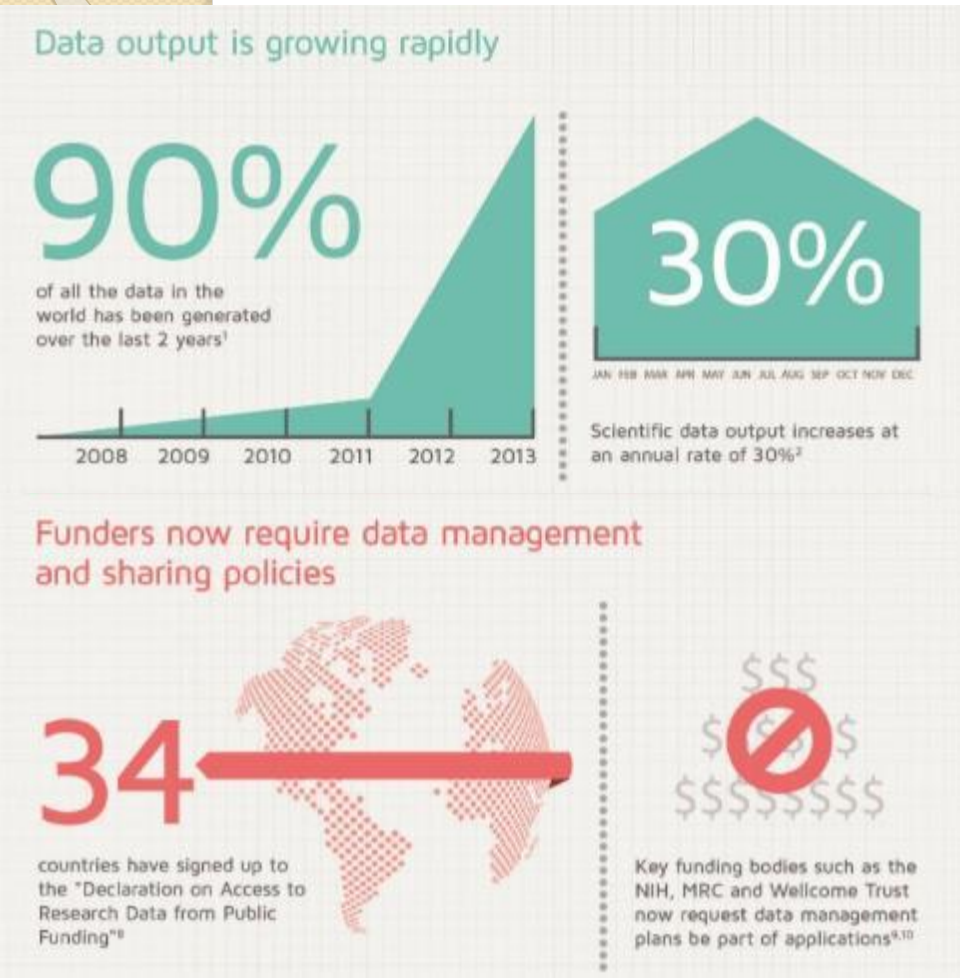
Pourquoi cet intérêt?

- Les données sont l'un des fondements des avancées scientifiques
- L'écologie est une discipline intégrative
- Explosion de données dans certains champs de l'écologie (« big data »)
- Certaines des questions posées à l'écologie ont une dimension continentale et/ou planétaire
- Développement des activités de synthèse scientifique

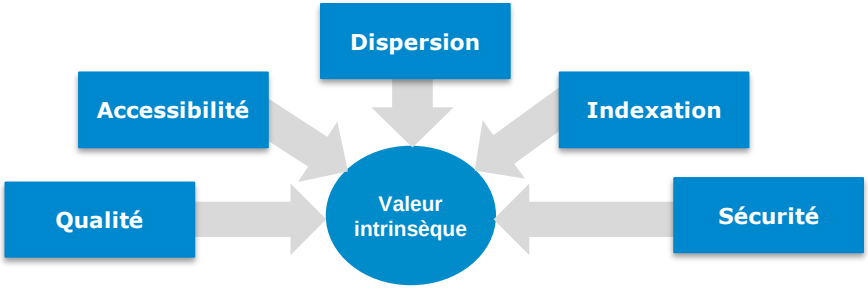


Lynch *et al.* (2015) *STOTEN* **534** : 173

Des données, mais…

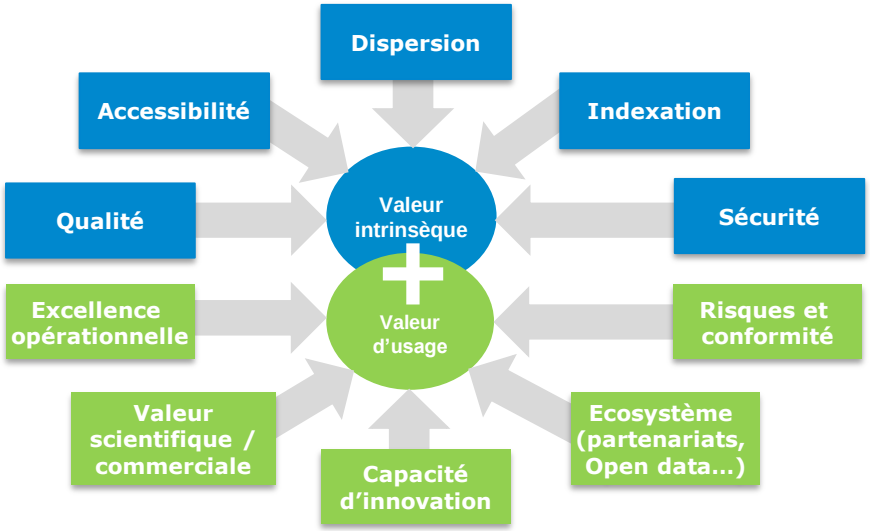


Deux perspectives sur les données



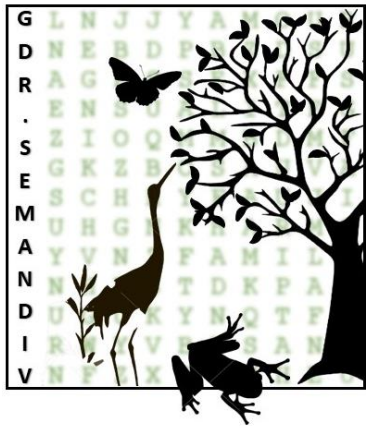
Enjeux techniques

Deux perspectives sur les données



Enjeux techniques

Enjeux métier

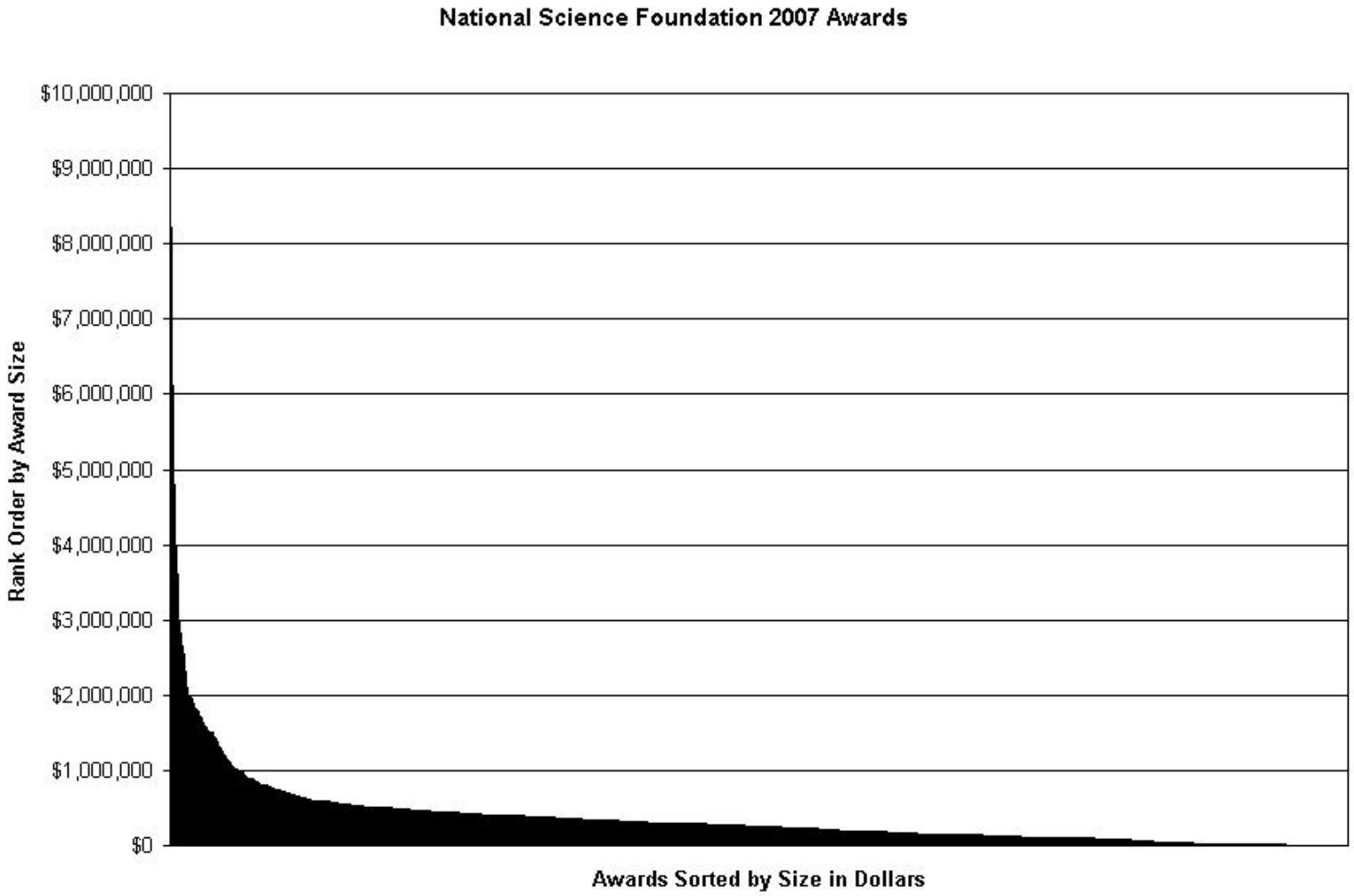


CARACTÉRISTIQUES DES DONNÉES EN ECOLOGIE

Les écologues génèrent une quantité considérable de données mais…

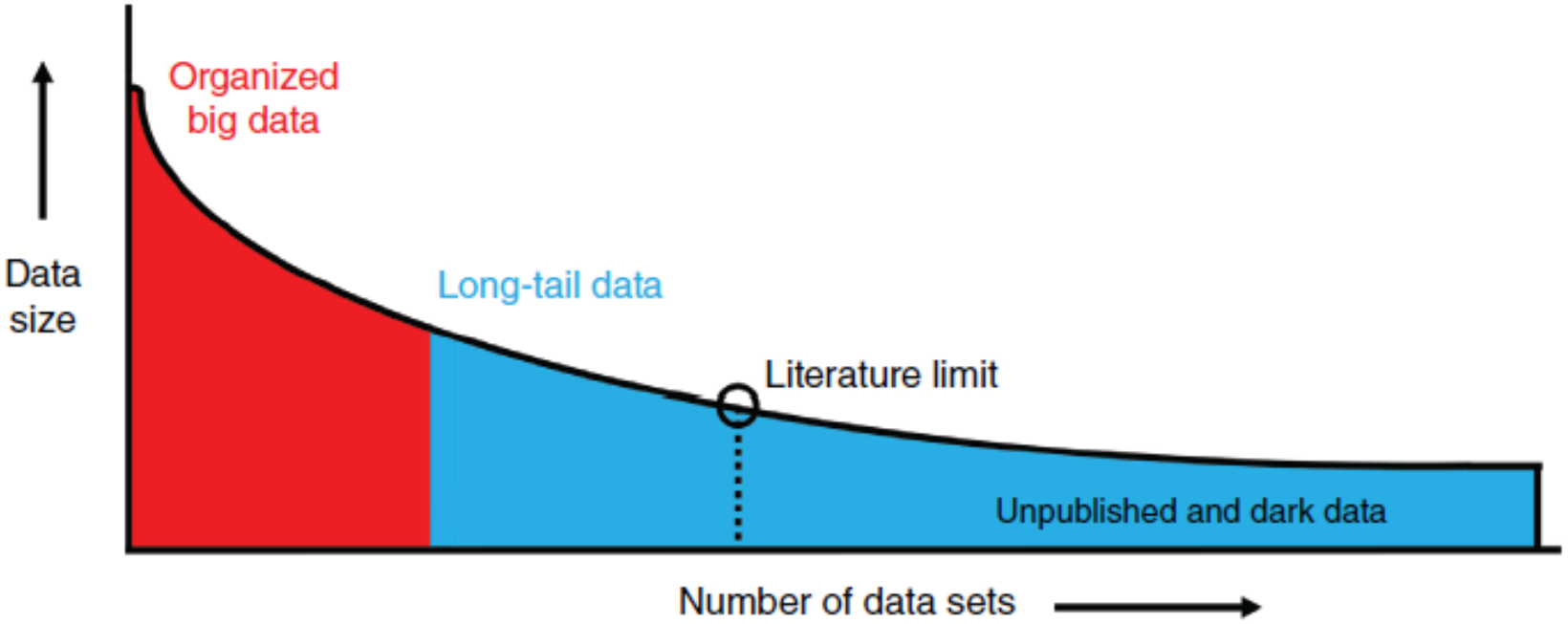
- majoritairement conduite par des individus et/ou groupes, sans réelle concertation
- peu de culture d'échange
- pas de procédures universellement admises et appliquées pour combiner et synthétiser ces données
- dominée par « long tail science » : beaucoup de petits jeux de données dispersés
  - hétérogénéité syntaxique
  - hétérogénéité sémantique

La structure du financement de la recherche…



Heidorn (2008) *Library Trends* **57**: 280

… et les « données obscures » de la « long tail science »



Ferguson *et al.* (2014) *Nature Neuro* **17** : 1142

"Donnée obscure": toute donnée qui n'est pas facilement découverte par un utilisateur potentiel (Heidorn, 2008)

Nature des jeux de données en écologie

Caractéristiques	Type de recherche	
	Coordonnée	Indépendante
Format	Homogène	Hétérogène
Collecte et saisie	Automatisées	Manuelles
Procédures	Uniformes	Contingentes
Archivage	Centralisé	Individuel
Dépôt	Bases disciplinaires et de référence	Equipes ou laboratoires
Entretien	Réalisé	Non réalisé
Accès	Ouvert	Obscur ou non protégé
Réutilisation	Immédiate	Rare
Valorisation	Dans le cadre de l'activité professionnelle	Souvent inaperçue

Heidorn (2008) *Library Trends* **57**: 280

Le gâchis des « données obscures »

- moins de 1% des données acquises en écologie sont accessibles après publication des résultats associés (Reichman *et al.* 2011) => opportunités perdues

- répétition inutile: coût humain et financier

- combinaison avec des données complémentaires pour répondre à des questions différentes et/ou plus larges impossible

INTEGRATION DES DONNEES POUR LA SYNTHESE SCIENTIFIQUE

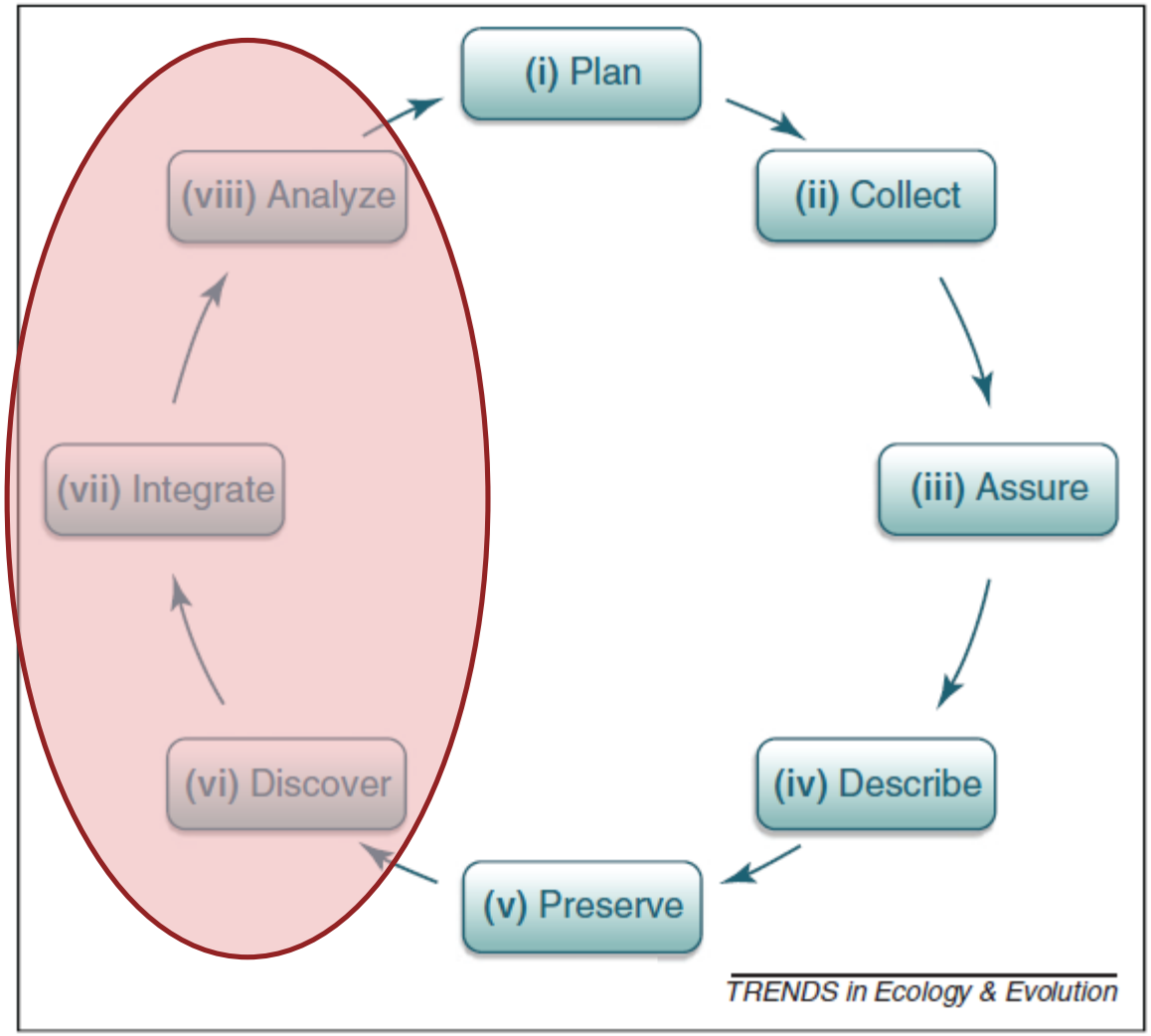
La synthèse scientifique

- Combinaison et intégration de différentes recherches afin d'améliorer la généralité et l'applicabilité des résultats de la recherche scientifique

- Elle peut se faire au sein d'une discipline, entre disciplines et entre secteurs d'activité professionnelle: elle n'est ***pas*** synonyme d'interdisciplinarité

Hampton & Parker (2011)
BioScience 61: 900

Synthèse et cycle de vie des données



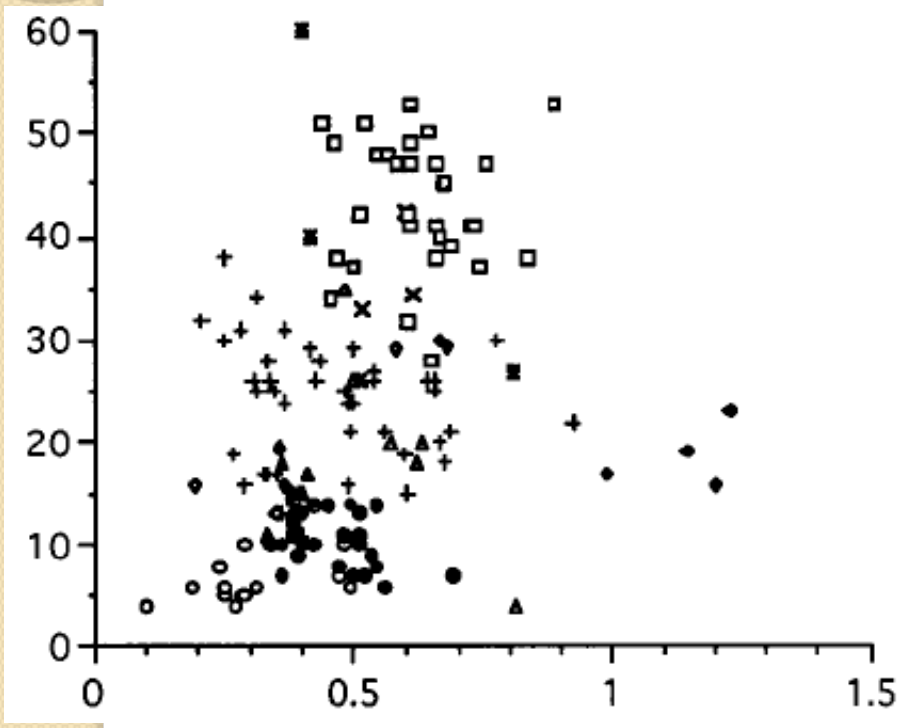
Michener & Jones (2012) *TREE 27* : 85

La synthèse scientifique en sciences de la biodiversité

- **Les questions relatives à la biodiversité sont généralement très complexes:**
  - Concepts pas toujours bien définis
  - La robustesse statistique et la généricité des résultats de la recherche en biodiversité et écologie doivent être améliorées
- **Le problème des petits jeux de données et/ou des données obscures**
  - Découverte et accessibilité
  - Hétérogénéité sémantique et syntaxique

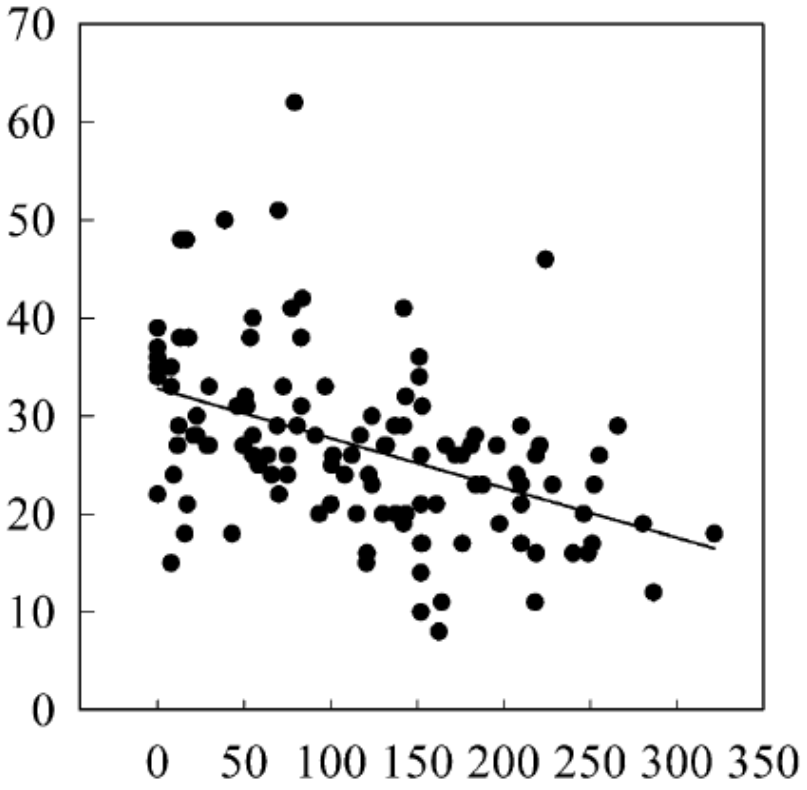
Ex. 1 : quelle est la relation entre diversité des plantes et fertilité des prairies permanentes ?

Nombre d'espèces de plantes



Teneur en azote total du sol (%)

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



Fertilisation azotée (kg N ha^{-1} an^{-1})

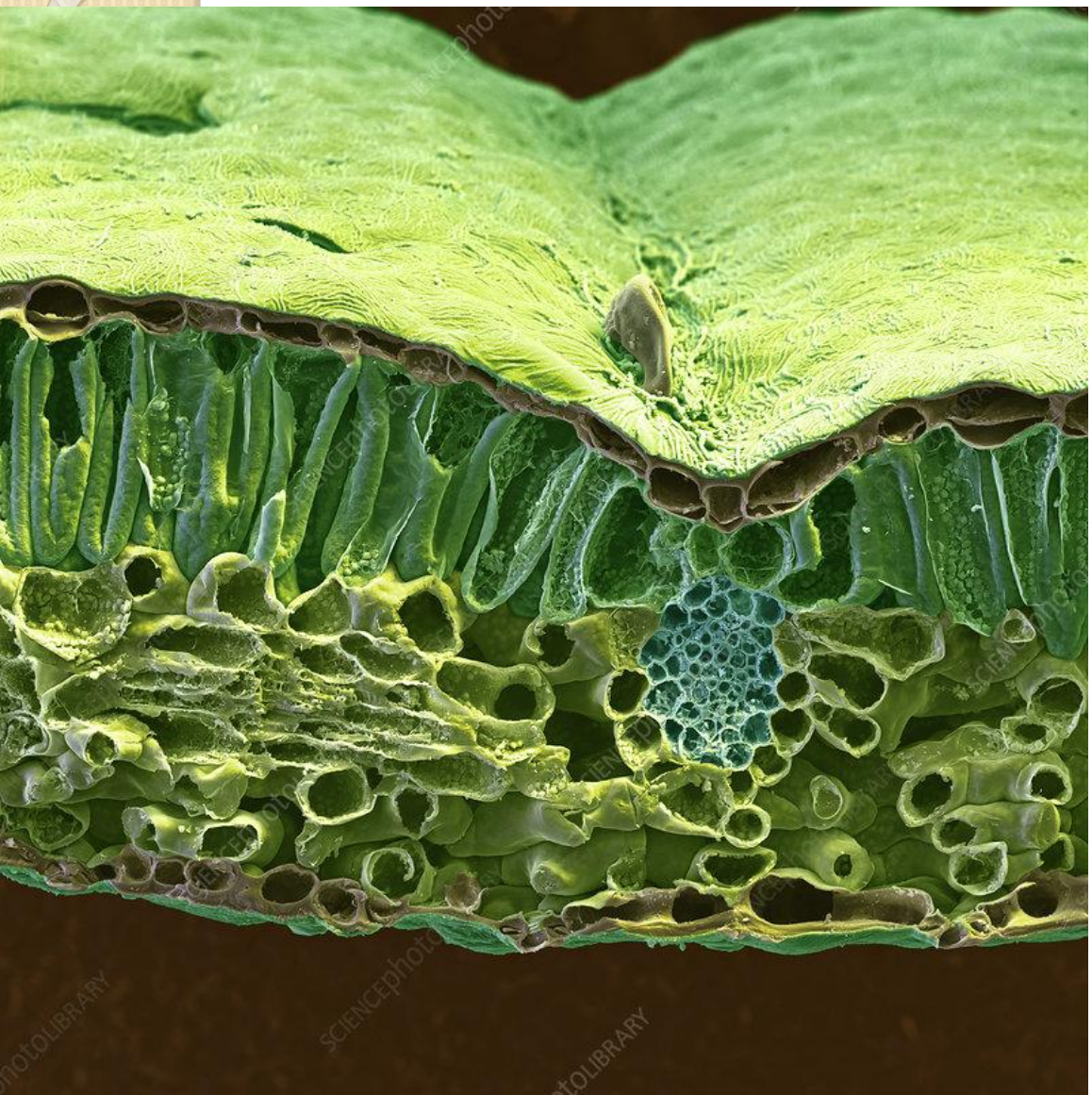
Klimek *et al.* (2007)
Biol Cons 134: 559

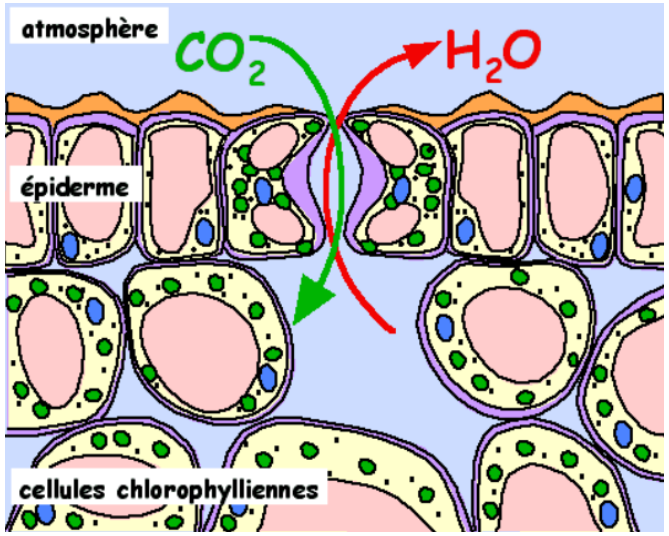
Les éléments nécessaires à la compréhension de cette relation

- **Définition des concepts:**
  - biodiversité
  - fertilité
- **Choix des variables descriptives:**
  - nombre d'espèces (indicateur de biodiversité)
  - composantes de la disponibilité en azote (indicateur de fertilité)
- **Méthodes**
- **Interactions avec les autres facteurs:**
  - contexte biogéographique
  - contexte pédo-climatique
  - autres ressources
  - autres pratiques de gestion
  - …

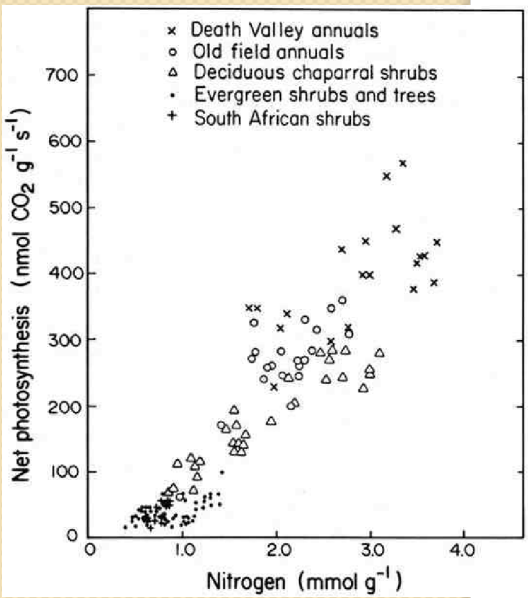
Ex. 2: traits et activité des feuilles

La feuille et la photosynthèse: porte d'entrée dans le monde vivant

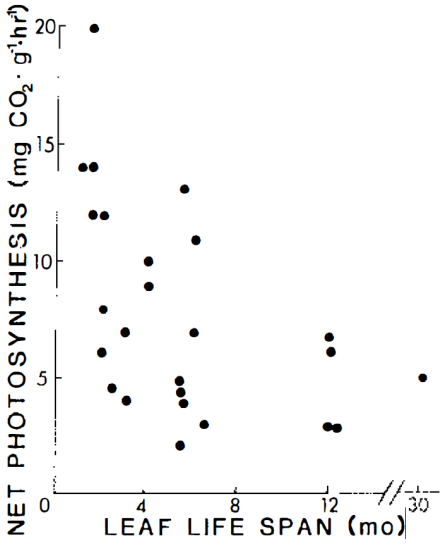




QUELS SONT LES CONTRÔLES SUR LA VITESSE DE PHOTOSYNTHÈSE ?

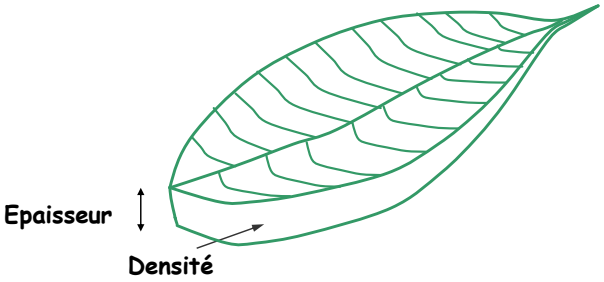


Field & Mooney (1986)
On the Economy of Plant Form



Chabot & Hicks (1982)
Ann Rev Ecol Syst 13: 229

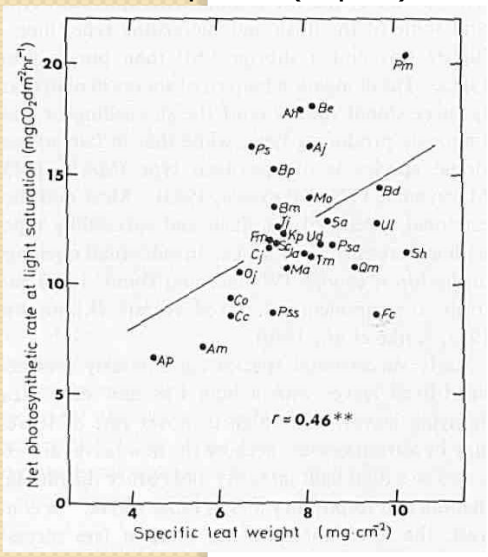
I – La structure morpho-anatomique des feuilles



SSF (SLA)= Surface d'une feuille/Masse sèche

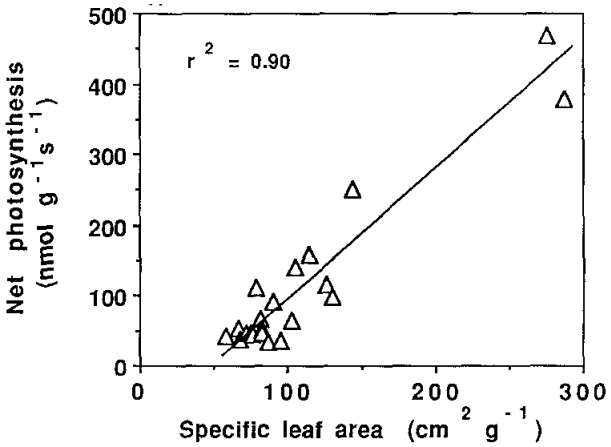
[MSF (LMA) = Masse sèche/surface]

30 espèces d'arbres de forêt tempérée (Japon)



Koike (1988)
Pl Sp Biol 3: 77

23 espèces d'arbres de la forêt amazonienne



Reich *et al.* (1991)
Oecologia 86: 16

29 espèces herbacées d'Europe centrale

Table 4. Leaf life span and leaf properties for a subsample of the species listed in Table 1 and 2

Plant species	Longevity (d)	LWA (g m^{-2})	Nitrogen content (% d.m.)	Nitrogen content (mmol m^{-2})	A_{sap}[a]	LWA/A_{sap}[b]
Low altitude:						
A.m.	42	51.8	3.88	193	17.9	2.89
R.a.	55	65.4	2.93	–	16.1	4.06
P.a.	57	52.3	3.09	116	20.2	2.59
T.r.	61	33.9	4.50	109	16.3	2.08
C.c.	61	37.9	3.50	95	19.0	1.99
C.l.	66	51.6	3.46	156	22.8	2.26
T.o.	66	34.4	3.60	81	22.7	1.51
C.p.	78	32.4	–	–	–	–
G.r.	87	52.6	2.96	91	12.5	4.21
P.e.	94	42.3	1.82	54	10.7	3.96
High altitude:						
P.v.	41	69.0	4.51	193	20.2	3.41
O.d.	48	25.4	5.00	129	16.4	1.55
D.c.	54	48.8	4.04	141	21.1	2.31
L.a.-p.	58	68.0	2.98	145	14.7	4.63
P.a.	61	64.5	2.27	136	20.2	3.19
G.m.	64	69.9	2.81	175	14.9	4.69
T.a.	66	36.5	2.94	77	10.6	3.44
A.e.	70	45.0	4.62	168	16.4	2.75
P.c.	73	60.6	3.28	168	16.4	3.70
P.g.	76	74.1	2.47	139	12.3	6.02
L.m.	77	79.4	3.62	200	23.9	3.32
E.u.	78	47.6	3.63	138	22.5	2.12
L.a.	80	62.5	2.36	111	11.0	5.68
G.r.	91	70.4	2.89	152	13.7	5.14
R.g.	93	64.5	3.37	173	19.4	3.32

[a] A_{sap} µmol m^{-2} s^{-1}; [b] LWA/A_{sap} g s µmol^{-1}

Statistical data:
Dependent variable: mean leaf life span (Spearman rank correlation)

Independent variable	*r*	R^2	*n*	*P*	sign.
LWA	0.09	0.01	25	0.336	n.s.
N (% d.m.)	−0.45	0.20	24	0.043	*
N (mmol m^{-2})	−0.17	0.03	23	0.228	n.s.
A_{sap}	−0.29	0.08	24	0.086	n.s.
LWA/A_{sap}	0.38	0.14	24	0.034	*

Linear regression: $y = a + b$ (life span), see Fig. 1

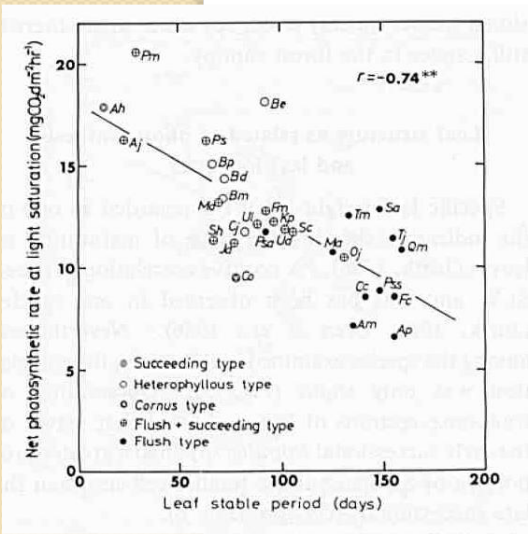
y	*a*	*b*	R^2	*P* of *b*	sign.
N (% d.m.)	5.23	−0.03	0.28	0.004	**
LWA/A_{sap}	1.05	0.03	0.17	0.020	*

* $P \leq 0.05$, ** $P \leq 0.01$, n.s. not significant

Diemer *et al.* (1992)
Oecologia 89: 10

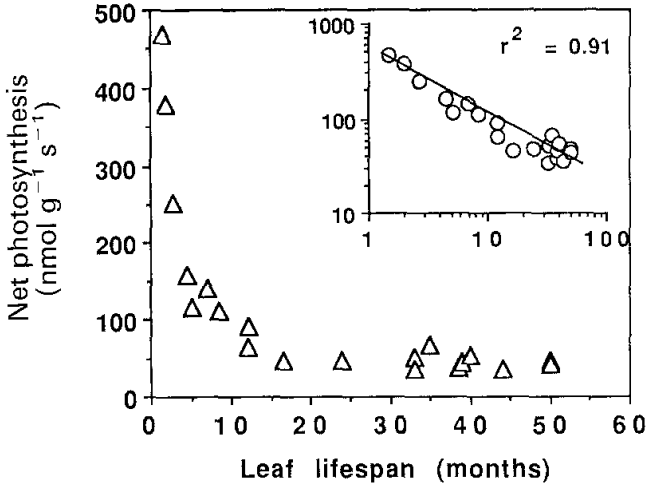
II – La durée de vie des feuilles

30 espèces d'arbres de forêt tempérée (Japon)



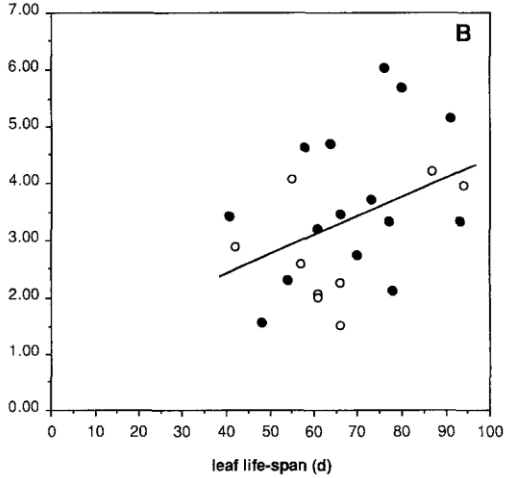
Koike (1988)
Pl Sp Biol 3: 77

23 espèces d'arbres de la forêt amazonienne



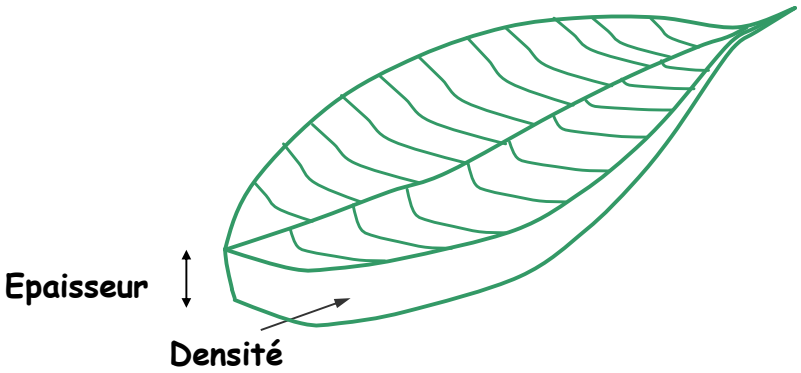
Reich *et al.* (1991)
Oecologia 86: 16

29 espèces herbacées d'Europe centrale



Diemer *et al.* (1992)
Oecologia 89: 10

Homogénéisation des variables pour la synthèse => travail sémantique (1)



SLA = surface/masse
= 1/LMA (SLW, LWA)

- Koike (1988) and Diemer *et al.* (1992) presented data on leaf mass per area, whereas Reich *et al.* (1991) provided data on the inverse, SLA. All data were converted to SLA to enable comparison with data from Reich *et al.* (1992), but otherwise, use of this measure, rather than its inverse, is arbitrary. Neither Koike (1988) nor Diemer *et al.* (1992) expressed *Amax* on a mass basis, but they did provide *Amax* values on an area basis, which I combined with SLA data to obtain estimated mass-based *Amax* rates.

Reich *et al.* (1993)
Functional Ecology 7: 721

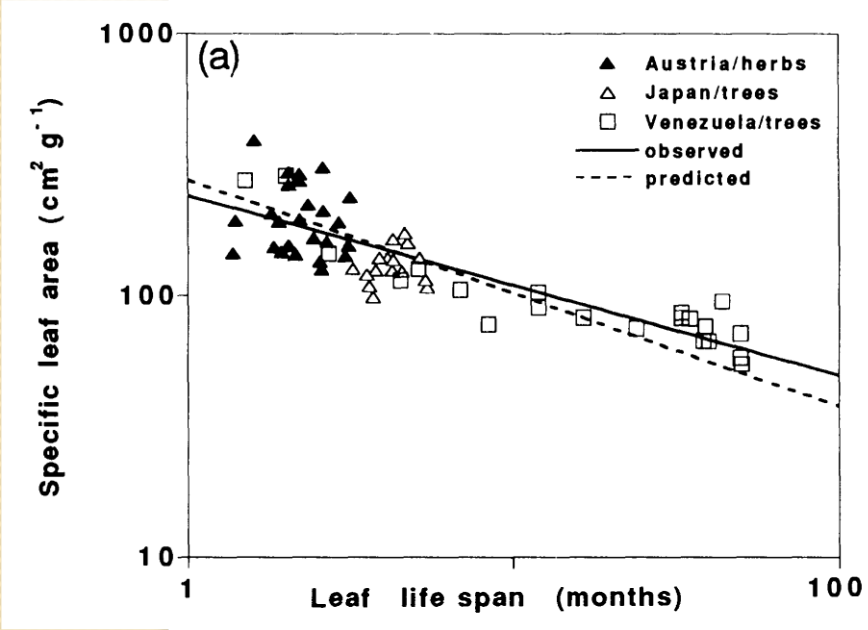
Homogénéisation des variables pour la synthèse => travail sémantique (2)

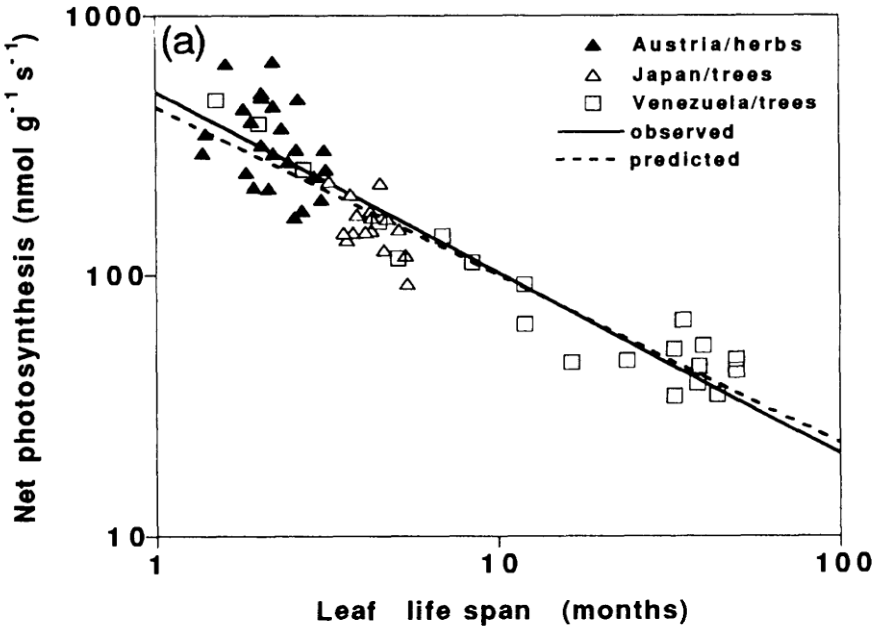
Leaf life span
≠
Leaf stable period
≠
Functional life span

- Diemer *et al.* (1992) and Reich *et al.* (1991) provided data on mean and median leaf life spans, respectively, but in the latter study, mean and median leaf life spans were similar. Diemer *et al.* (1992) used a measure of the 'functional leaf life span' that did not include the time required for the final decay of leaves during senescence prior to shedding. Koike (1988) provided both a measure of the 'leaf stable period' and a measure of the mean leaf life span. The 'leaf stable period' was defined to include only the period when *Amax* was stable and maximal. Thus, the 'leaf stable period' is less than the total leaf life span, and also often less than the 'functional life span' used by Diemer *et al.* (1992). Therefore, mean life-span data (available for 16 species, Koike 1988) were used.

Reich *et al.* (1993)
Functional Ecology 7: 721

Test sur les jeux de données combinés



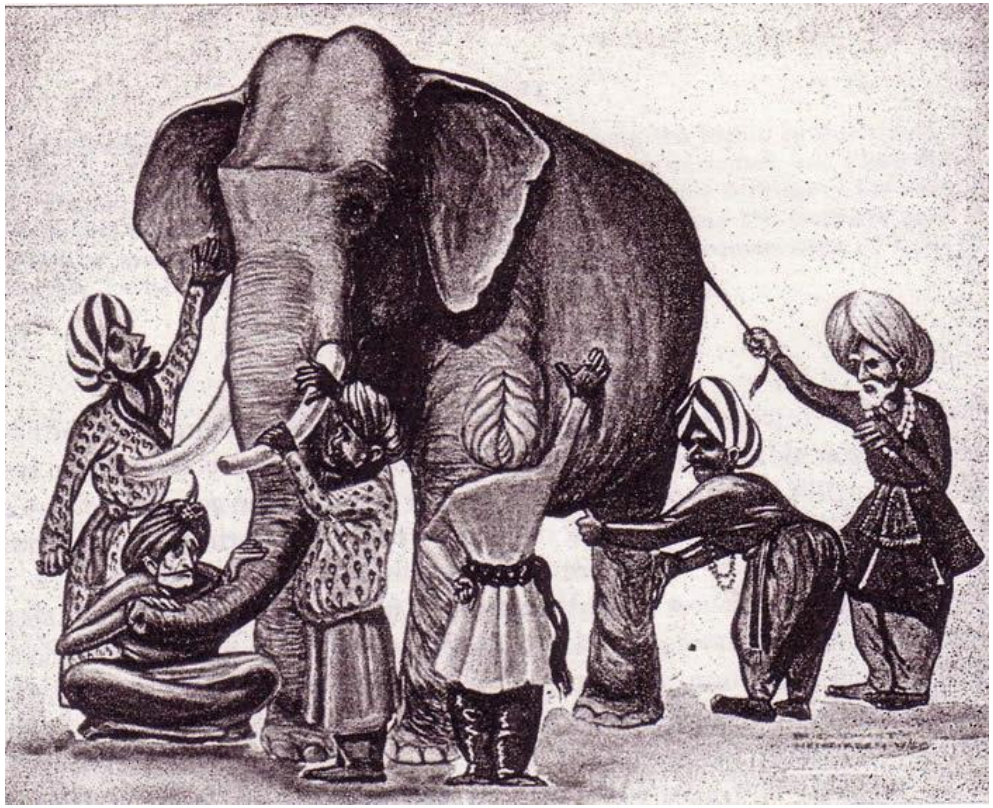


Reich *et al.* (1993)
Functional Ecology 7: 721

Functional Ecology 1993
7, 721–725

Reconciling apparent discrepancies among studies relating life span, structure and function of leaves in contrasting plant life forms and climates: ‘the blind men and the elephant retold’

P. B. REICH



La parabole des aveugles (*i.e.* les scientifiques) et de l’éléphant (*i.e.* les lois générales)

Le réseau « Global Plant Trait Network »: 40 chercheurs, 15 pays

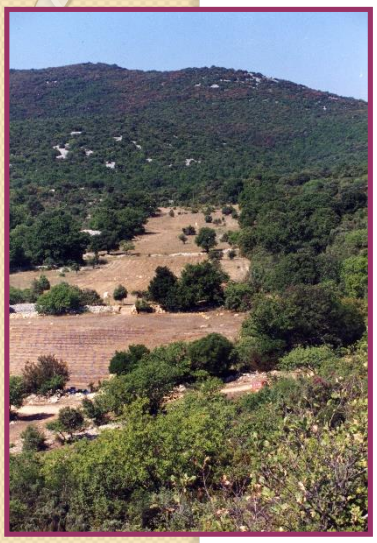
Fonction d'acquisition du carbone:

- vitesses de photosynthèse max. et de respiration
- surface spécifique et durée de vie
- composition chimique: N et P

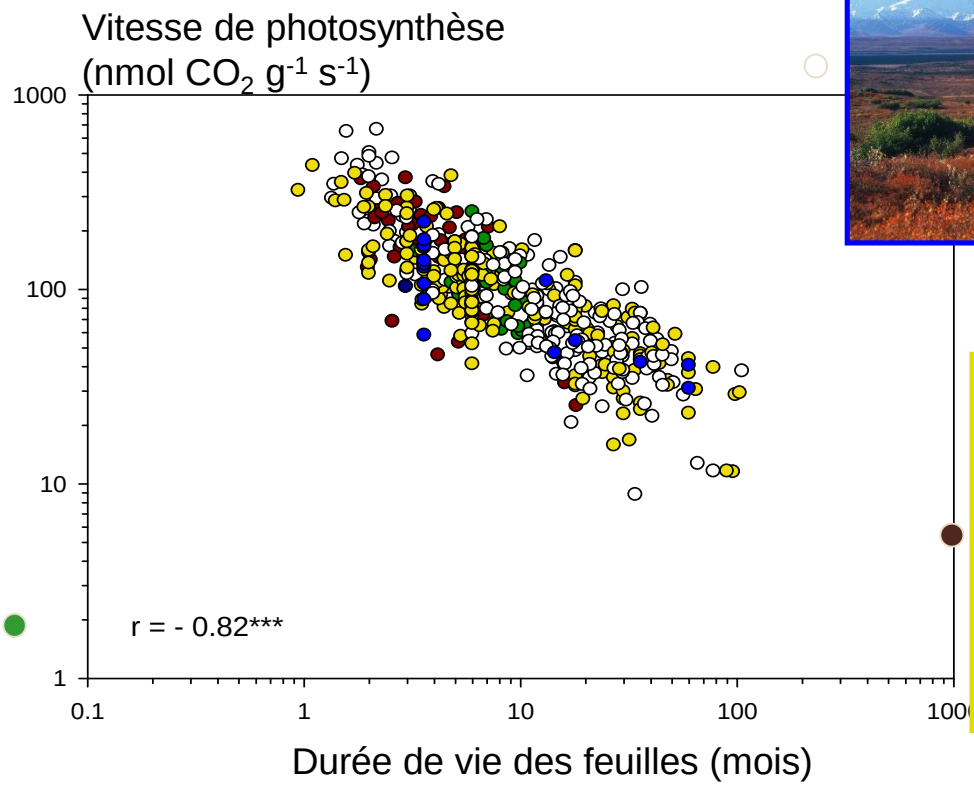
***Base de données mondiale*:**

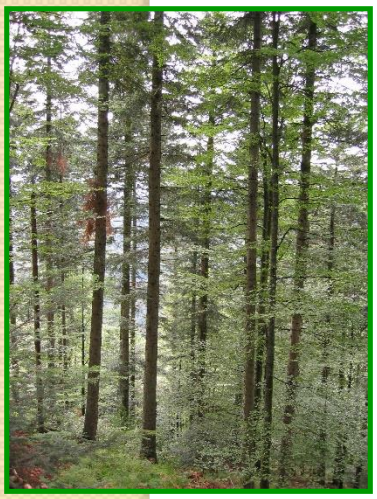
- 2548 espèces (~1% du nombre total d'espèces de végétaux terrestres);
- 219 familles; aucune restriction sur le type d'espèce (fougères, herbacées, arbustes, arbres)

Un schéma universel de fonctionnement des végétaux (1)



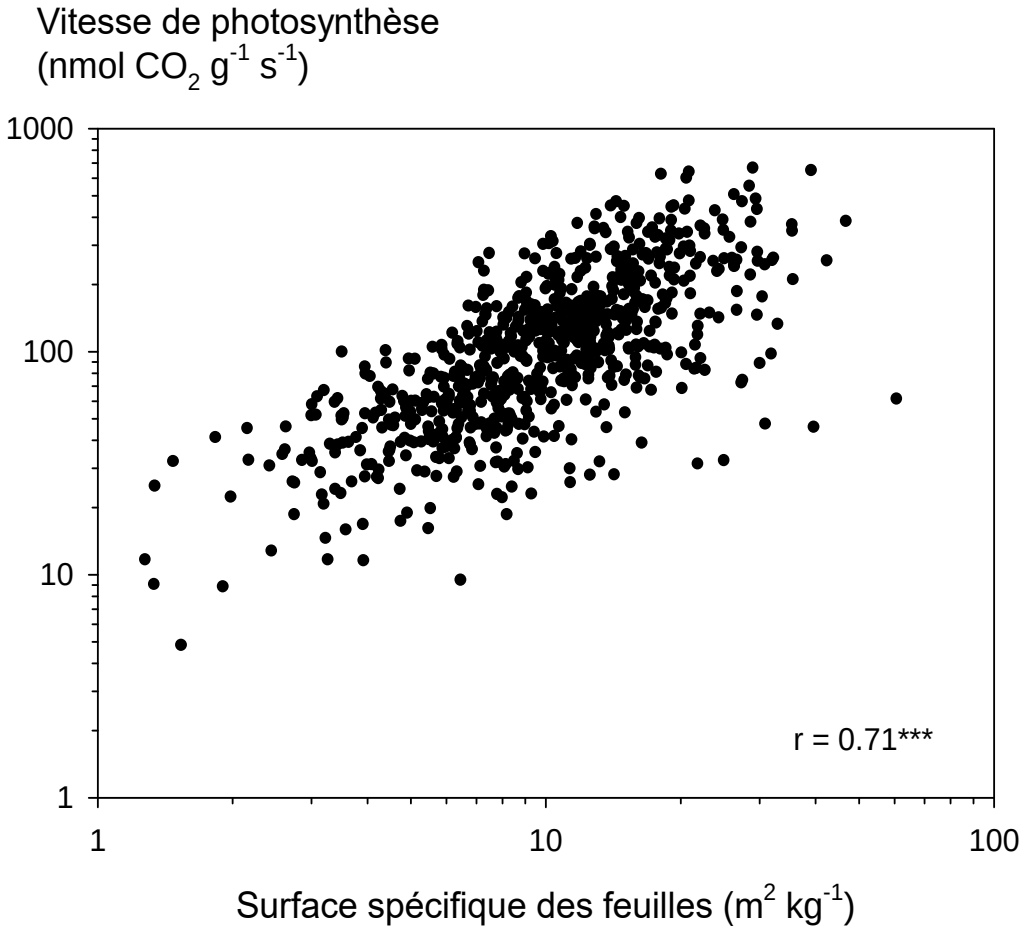
- Autres biomes





Wright *et al.* (2004)
Nature, 428: 821

Un schéma universel de fonctionnement des végétaux (2)

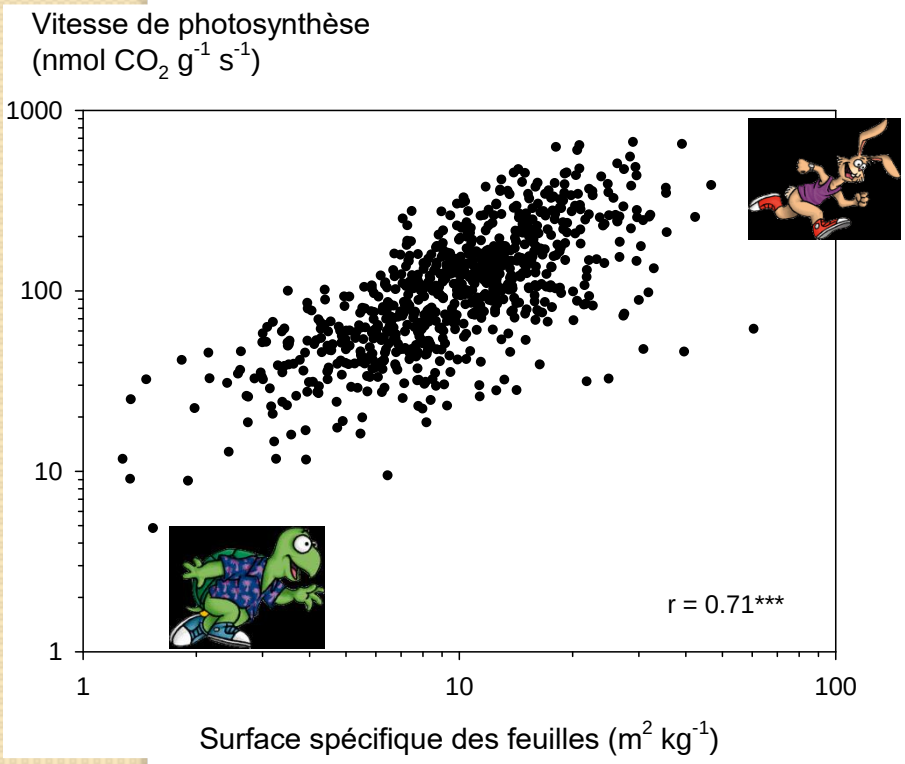


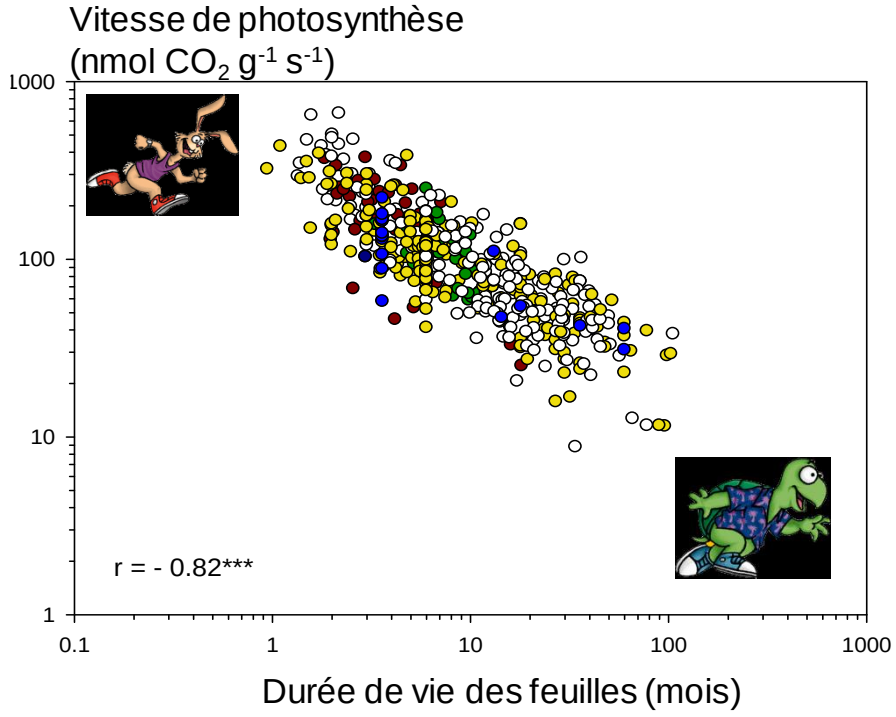
Wright *et al.* (2004)
Nature, 428: 821

The worldwide leaf economics spectrum

Ian J. Wright[1], Peter B. Reich[2], Mark Westoby[1], David D. Ackerly[3], Zdravko Baruch[4], Frans Bongers[5], Jeannine Cavender-Bares[6], Terry Chapin[7], Johannes H. C. Cornelissen[8], Matthias Diemer[9], Jaume Flexas[10], Eric Garnier[11], Philip K. Groom[12], Javier Gulias[10], Kouki Hikosaka[13], Byron B. Lamont[12], Tali Lee[14], William Lee[15], Christopher Lusk[16], Jeremy J. Midgley[17], Marie-Laure Navas[11], Ülo Niinemets[18], Jacek Oleksyn[2,19], Noriyuki Osada[20], Hendrik Poorter[21], Pieter Poot[22], Lynda Prior[23], Vladimir I. Pyankov[24], Catherine Roumet[11], Sean C. Thomas[25], Mark G. Tjoelker[26], Erik J. Veneklaas[22] & Rafael Villar[27]

NATURE | VOL 428 | 22 APRIL 2004 | www.nature.com/nature





Les leçons à tirer de cette course d'obstacles

- Les données doivent pouvoir être découvertes
- Leur réutilisation (par la personne qui les a collectées ou par des tiers) nécessite une mise en forme soignée et explicite
- C'est une histoire qui finit (plutôt) bien: mise en évidence d'une loi fondamentale du fonctionnement des plantes
- Mais… que d'efforts !

Vers une standardisation…

CSIRO PUBLISHING
www.publish.csiro.au/journals/ajb
Australian Journal of Botany, 2003, **51**, 335–380

A handbook of protocols for standardised and easy measurement of plant functional traits worldwide

J. H. C. Cornelissen[A,J], *S. Lavorel*[B], *E. Garnier*[B], *S. Díaz*[C], *N. Buchmann*[D], *D. E. Gurvich*[C], *P. B. Reich*[E], *H. ter Steege*[F], *H. D. Morgan*[G], *M. G. A. van der Heijden*[A], *J. G. Pausas*[H] *and H. Poorter*[I]

Global Change Biology
Global Change Biology (2011) **17**, 2905–2935, doi: 10.1111/j.1365-2486.2011.02451.x

TRY – a global database of plant traits

J. KATTGE*, S. DÍAZ†, S. LAVOREL‡, I. C. PRENTICE§, P. LEADLEY¶, G. BÖNISCH*, E. GARNIER‖, M. WESTOBY§, P. B. REICH**, ††, I. J. WRIGHT§, J. H. C. CORNELISSEN‡‡,

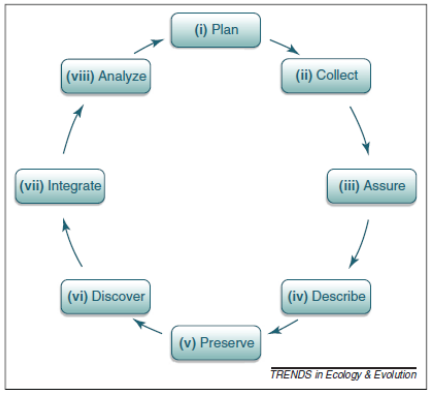
Journal of Ecology
Journal of Ecology 2017, **105**, 298–309
doi: 10.1111/1365-2745.12698

Towards a thesaurus of plant characteristics: an ecological contribution

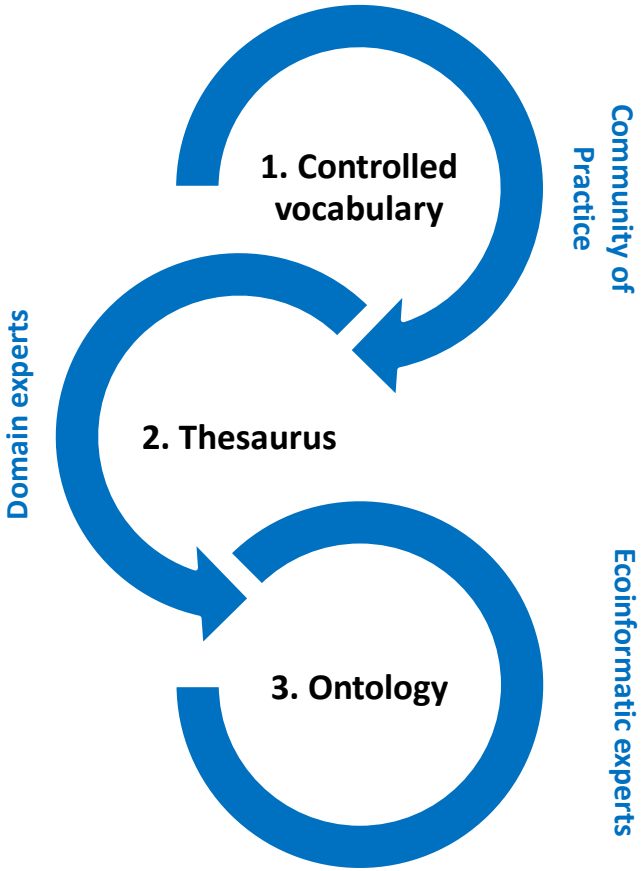
Eric Garnier*[,1,2], Ulrike Stahl[3,4,5], Marie-Angélique Laporte[1,4,6], Jens Kattge[3,4], Isabelle Mougenot[7], Ingolf Kühn[4,5,8], Baptiste Laporte[2], Bernard Amiaud[9,10], Farshid S. Ahrestani[11,12], Gerhard Bönisch[3], Daniel E. Bunker[13], J. Hans C. Cornelissen[14], Sandra Díaz[15], Brian J. Enquist[16], Sophie Gachet[17], Pedro Jaureguiberry[15], Michael Kleyer[18], Sandra Lavorel[19], Lutz Maicher[20,21], Natalia Pérez-Harguindeguy[15], Hendrik Poorter[22], Mark Schildhauer[23], Bill Shipley[24], Cyrille Violle[1], Evan Weiher[25], Christian Wirth[4,26], Ian J. Wright[27] and Stefan Klotz[5]

Plus de détails demain!

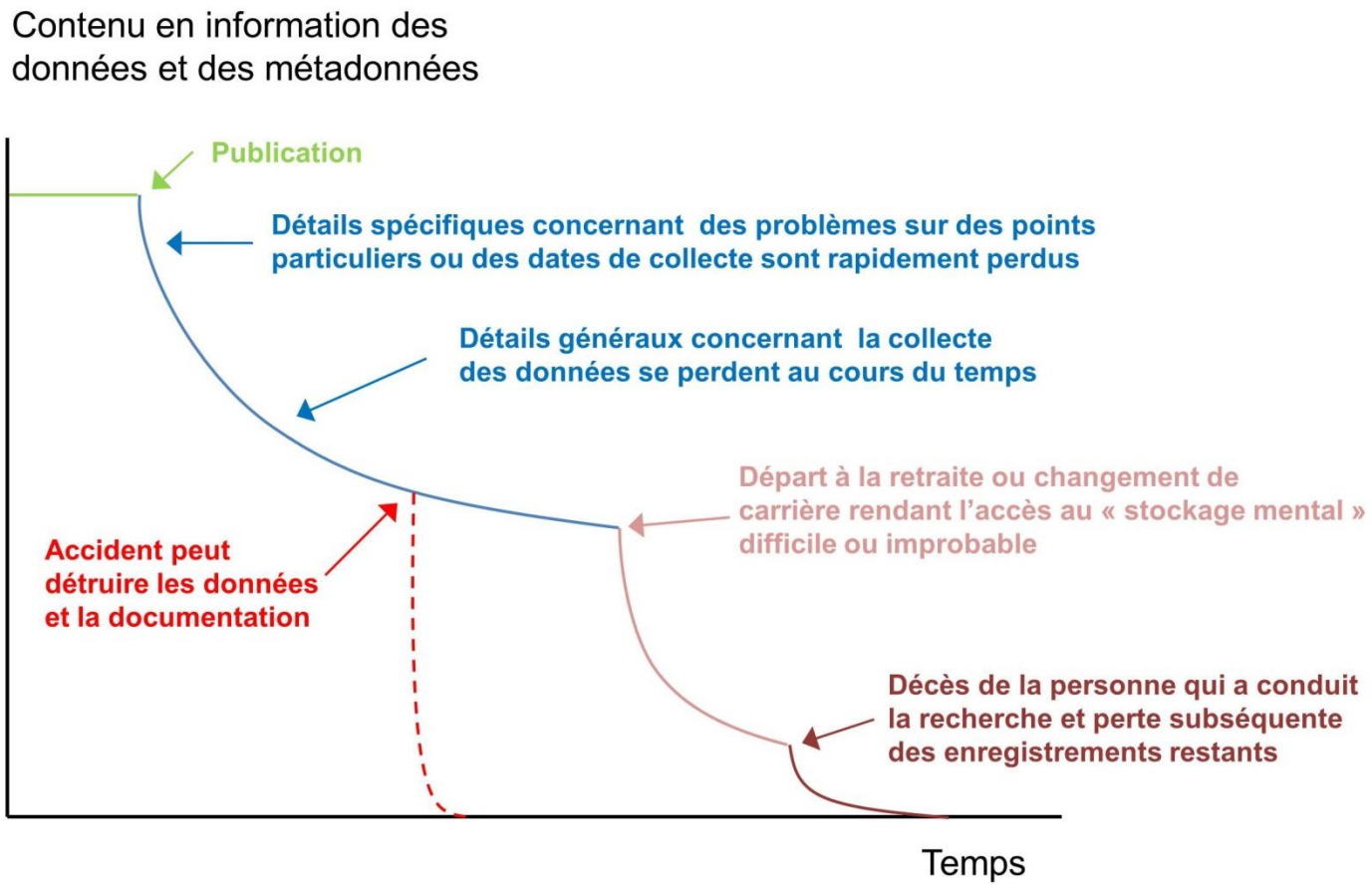

MERCI POUR VOTRE ATTENTION
DES QUESTIONS?



Cycle de vie des données: (VII) intégrer: sémantique

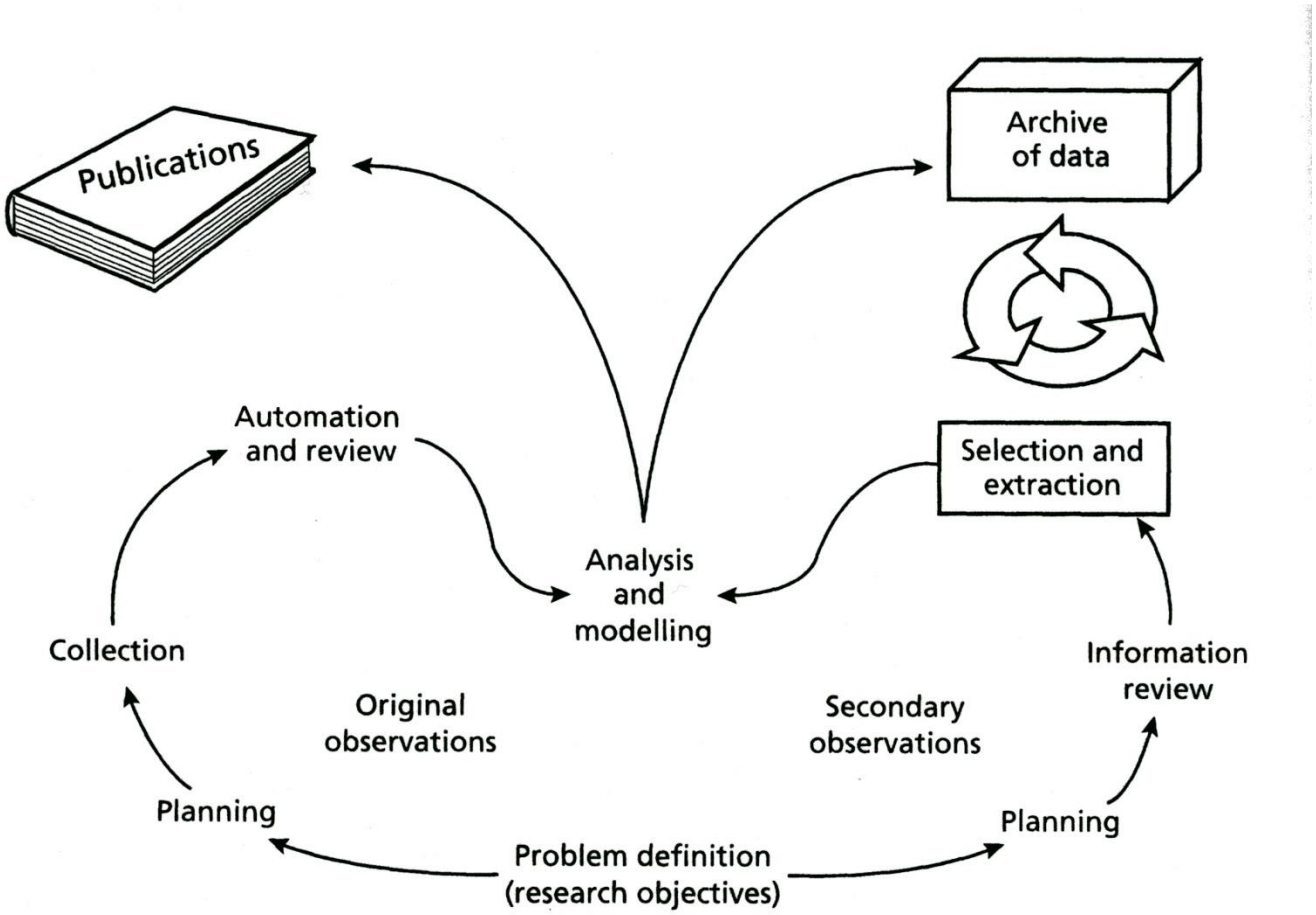


Données et information: une histoire d'entropie



Michener *et al.* (1997) *Ecol Appl 7* : 330

Les données et la recherche



Olson & McCord (2000)
Ecological Data

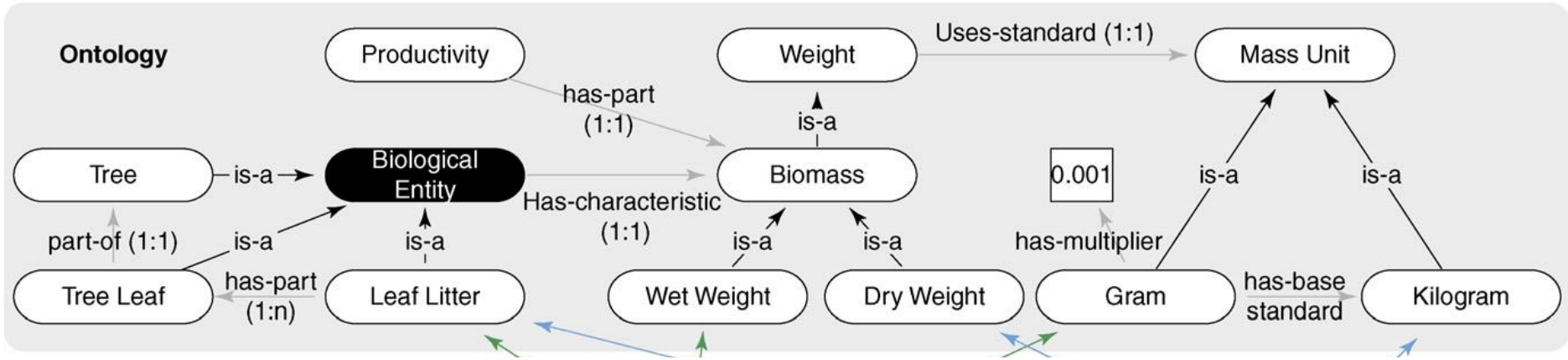
Intégrer: les ontologies

- En sciences de l'information, une ontologie est un mode formel d'organisation de la connaissance dans lequel:
  - chaque élément (ou concept) est défini precisément
  - chaque relation potentielle entre ces éléments est paramétrée ou contrainte (e.g. is_a, part_of, has_member, has_characteristic…)

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

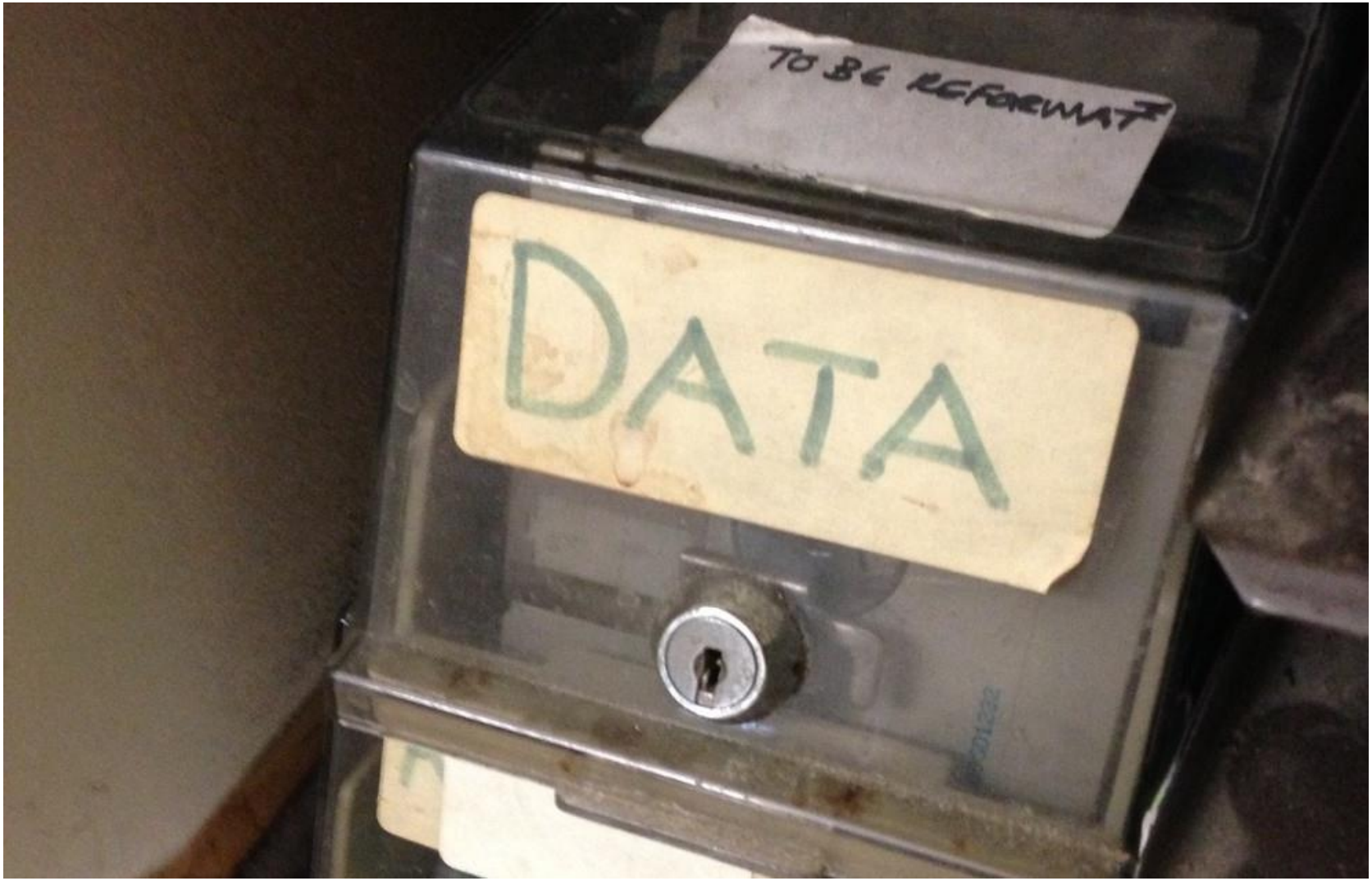
- Représentation explicite d'un domaine permettant d'effectuer des tâches raisonnées de façon automatique

Intégrer: exemple d'ontologie en écologie

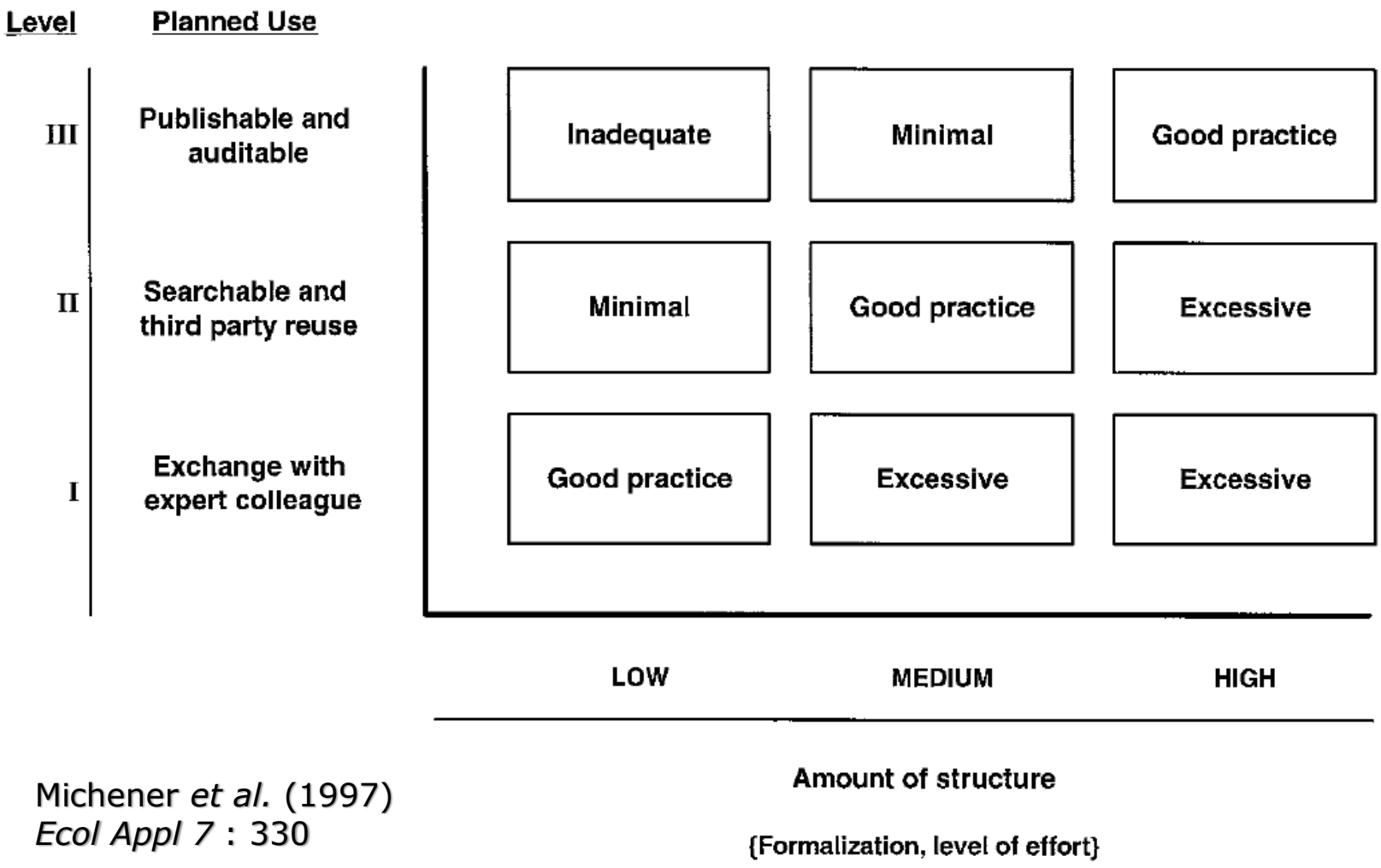


Madin *et al.* (2008) *TREE* 23: 159

Pour éviter ça…



Adéquation entre utilisation et effort de structuration



Michener *et al.* (1997)
Ecol Appl 7 : 330

En guise de conclusion (2)

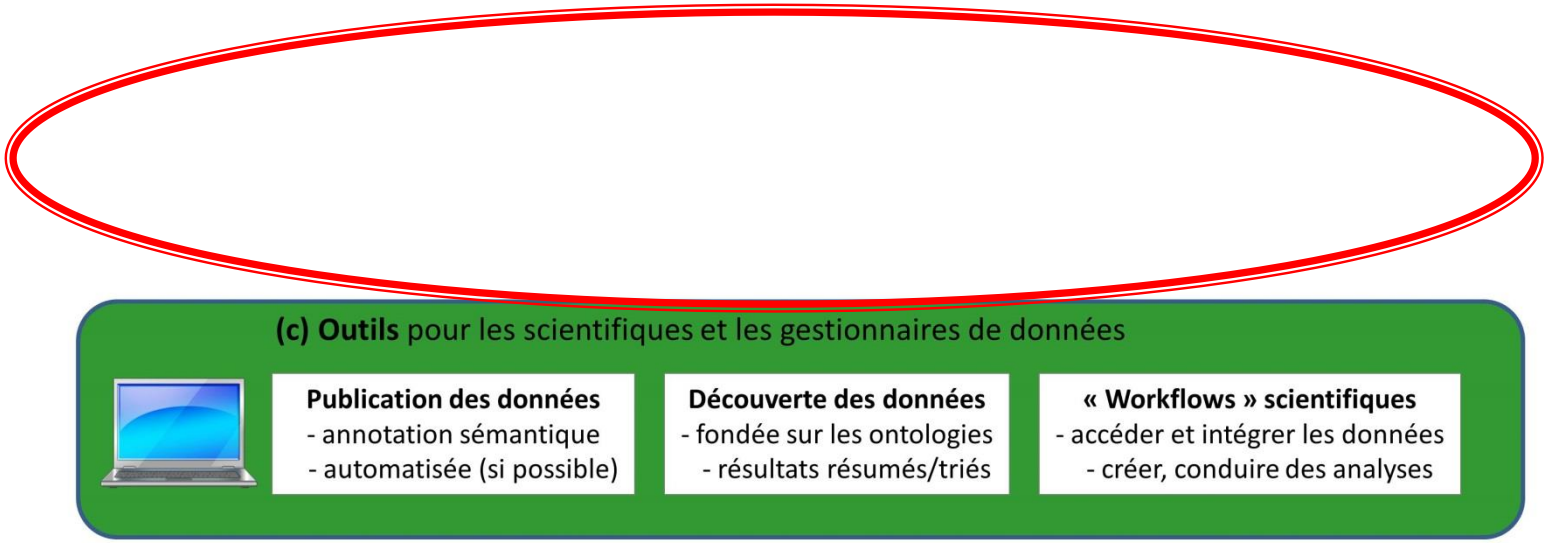
- Les données collectées peuvent être utilisées pour des questions qui ne sont pas encore formulées
- Pensez plus loin que votre propre étude pour que vos données puissent être réutilisées
- En route pour le « big data » en écologie… mais pas seulement!

L'émergence de l'écoinformatique:

- champ de recherche et de développement situé aux interfaces de l'écologie, de l'informatique et des sciences de l'information…
  (Jones *et al.* 2006, *Ann Rev Ecol Syst*)

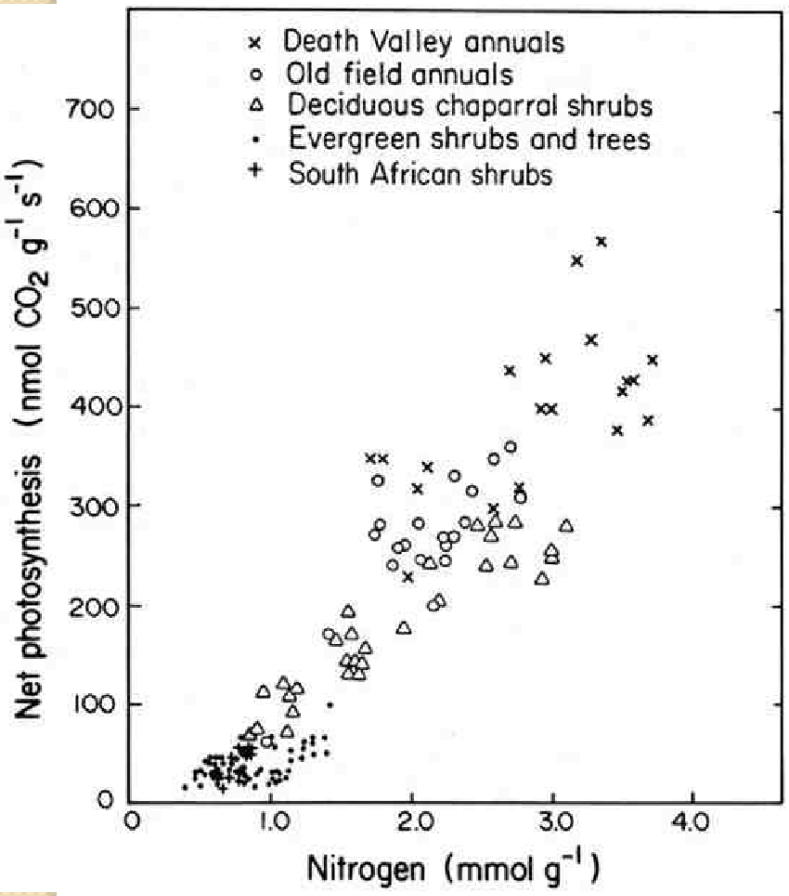
- … qui a pour objectif de permettre aux scientifiques de générer de nouvelles connaissances grâce à des outils et approches innovants pour découvrir, gérer, intégrer, analyser, visualiser et sauvegarder les données et les informations biologiques, environnementales et socioéconomiques pertinentes
  (Michener & Jones 2012, *TREE*).

Les objectifs de l'écoinformatique

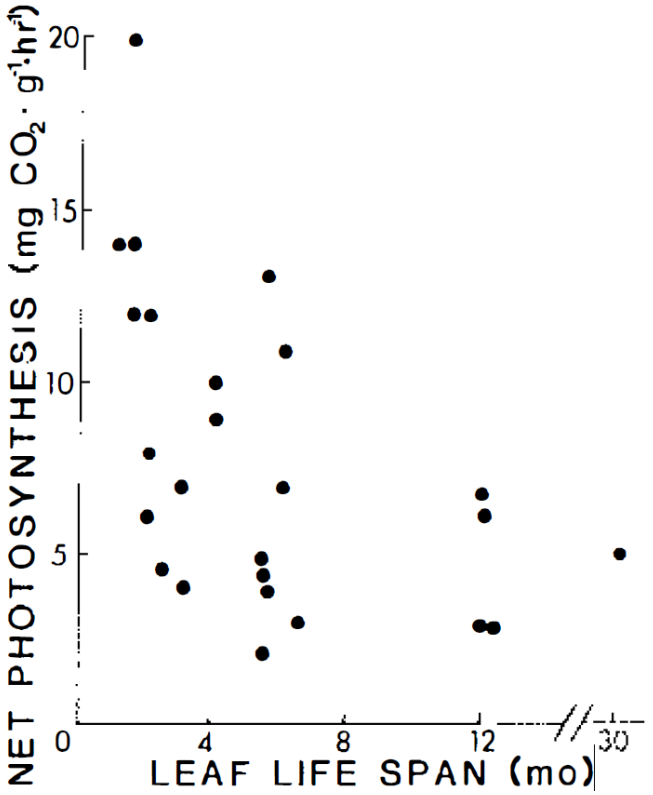


Madin *et al.* (2008) *TREE* 23: 159

Quels sont les contrôles sur la vitesse de photosynthèse?



Field & Mooney (1986)
On the Economy of Plant Form



Chabot & Hicks (1982)
Ann Rev Ecol Syst 13: 229