Formation GBIF sur la qualité, la publication et l'utilisation des données sur la biodiversité – Antananarivo, 04 – 05 avril 2016

Introduction à la qualité des données et à l'adéquation à l'usage

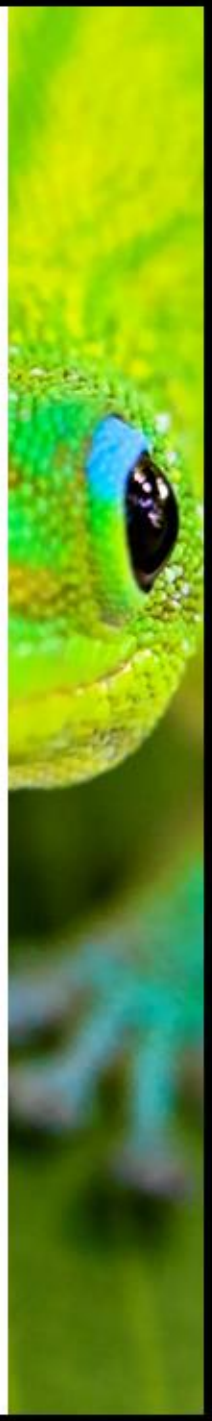
GBIF France (gbif@gbif.fr)

Présentation réalisée en collaboration avec Nicolas Noé
Développeur – Plateforme Belge Biodiversité
Global Biodiversity Information Facility (GBIF)

Pourquoi publier les données ?

21ème siècle = « siècle des données »

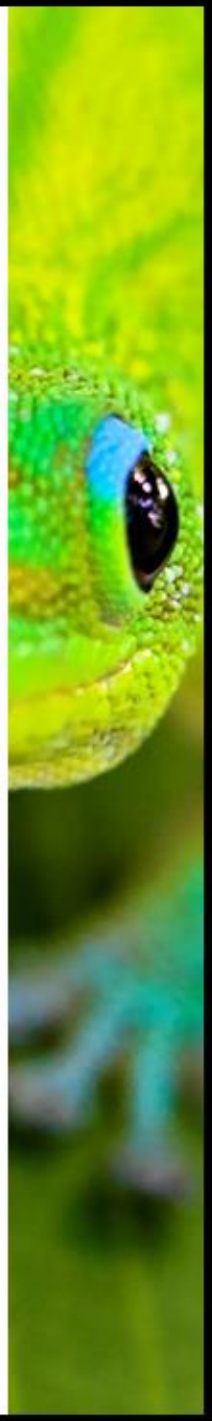
- La quantité de données augmente exponentiellement
- Le GBIF est un acteur de ce mouvement !
- Ces données **ont le potentiel** d'améliorer grandement nos connaissances et aptitudes



Des données à la compréhension…

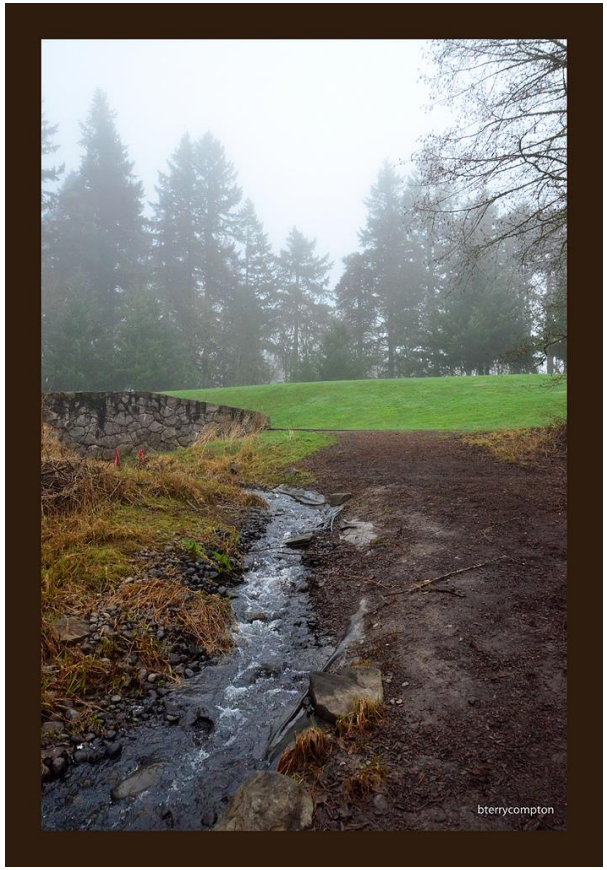
Des océans de données…

…des rivières d’informations…

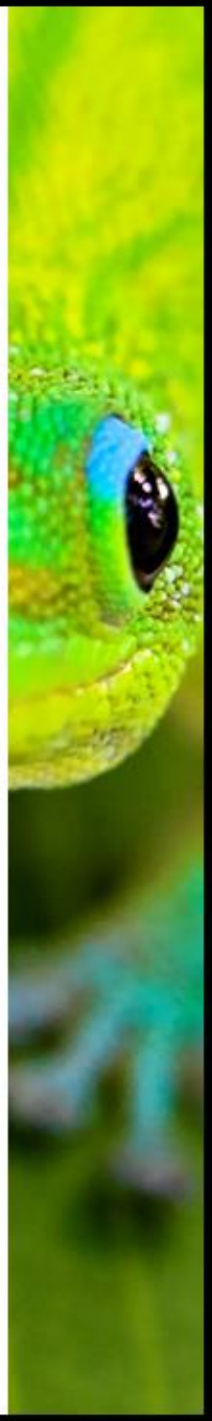


GBIF
www.gbif.org

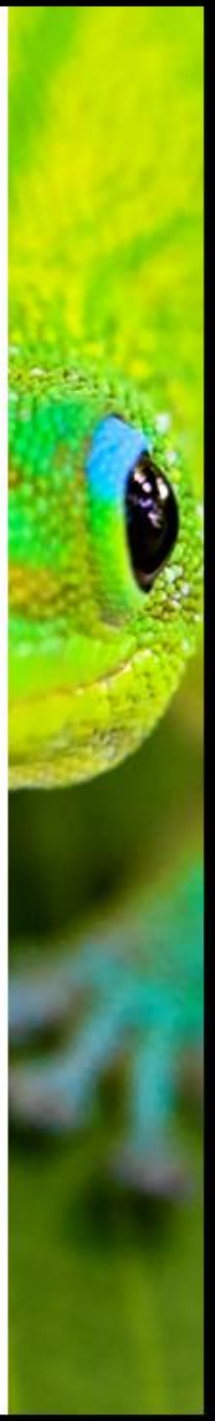
gbif France



… des ruisseaux de connaissances …



…des gouttes de compréhension



GBIF
www.gbif.org

gbif France

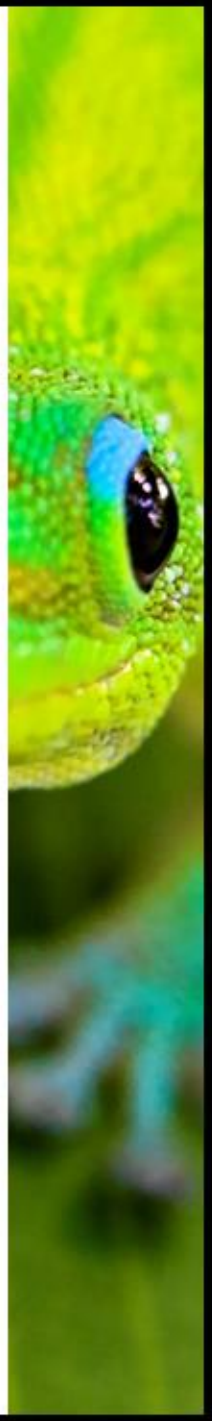
Usage des données de biodiversité

Recherches taxonomiques, modélisation/prédiction de la distribution des espèces, espèces invasives, dégradation des habitats, relations interspécifiques, ...

Mais aussi...

Organisation de la conservation, gestion de l'eau, éco-tourisme, histoire des sciences, chasse et pêche, rapatriement des données, ...

D'après Chapman, 2006



GBIF
www.gbif.org

gbif France

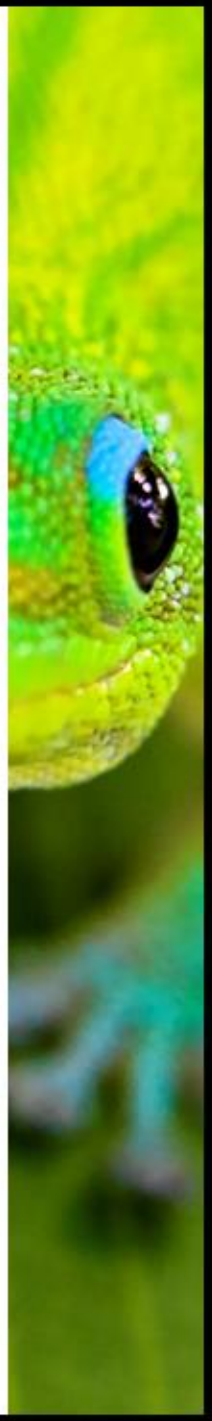
Adéquation à l'usage - définition

« Fitness-for-use »

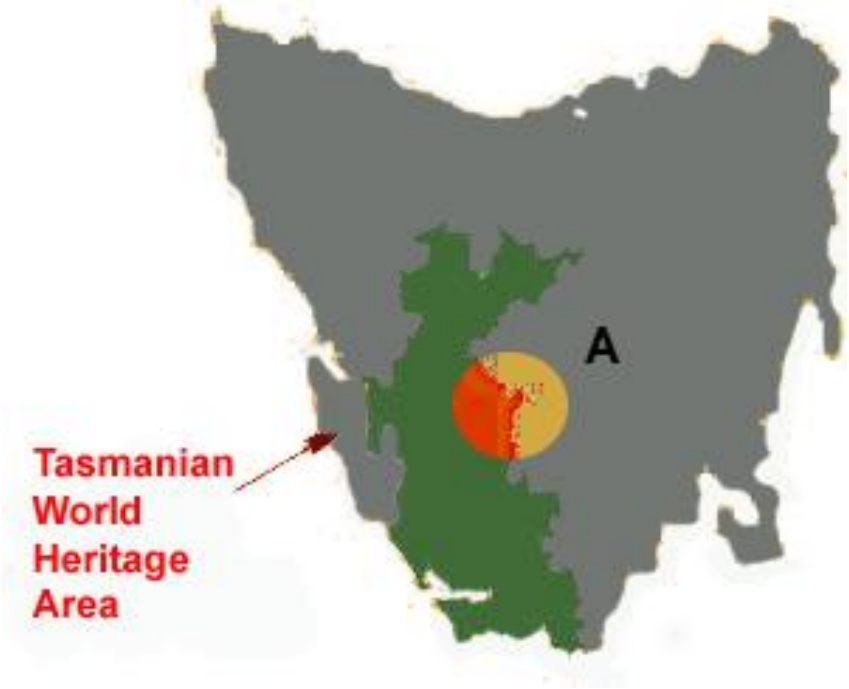
La qualité des données est un concept relatif qui dépend de l'usage qui est fait de ces données…

"The general intent of describing the quality of a particular dataset or record is to describe the fitness of that dataset or record for a particular use that one may have in mind for the data."

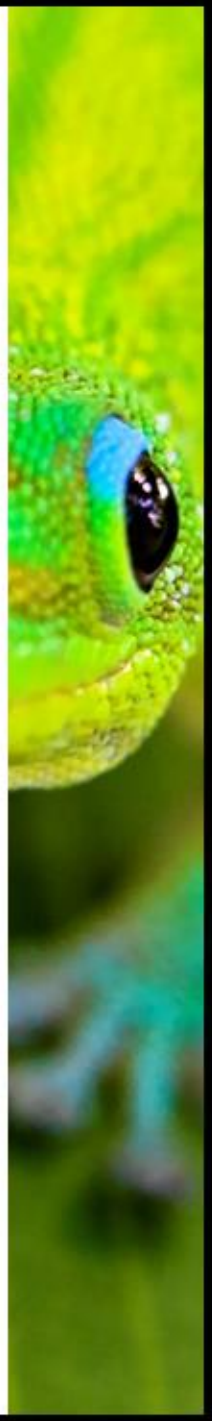
Chrisman, 1991



Adéquation à l’usage - exemple



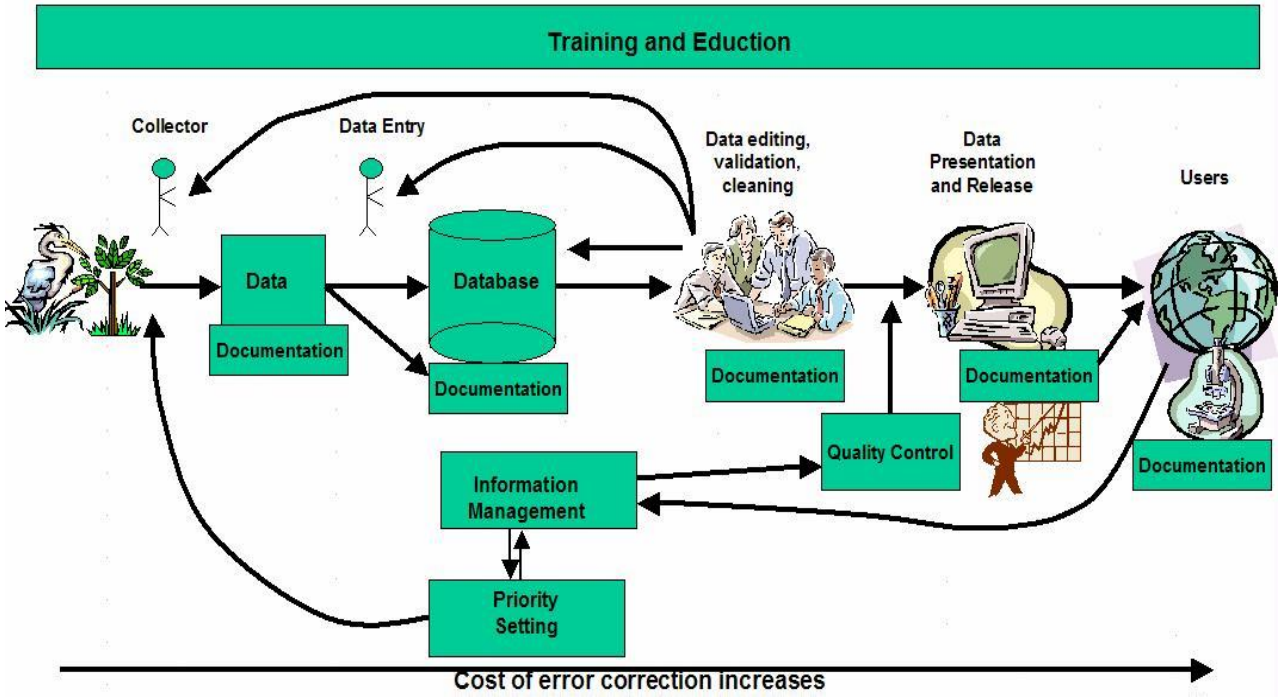
L’espèce est-elle présente en Tasmanie ?
L’espèce est-elle présente dans la réserve ?



GBIF
www.gbif.org

gbif France

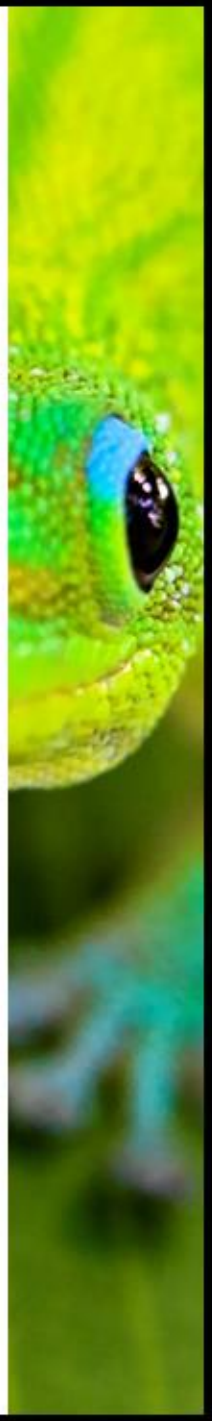
Chaîne des données et qualité



La perte de qualité survient à chaque étape.

La responsabilité en terme de qualité de données doit être assignée le plus tôt possible dans cette chaîne.

Chaîne des données et qualité

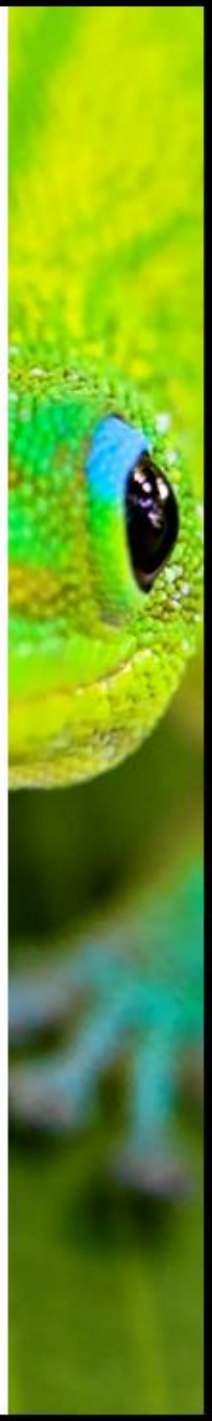


Chaque institution devrait avoir :

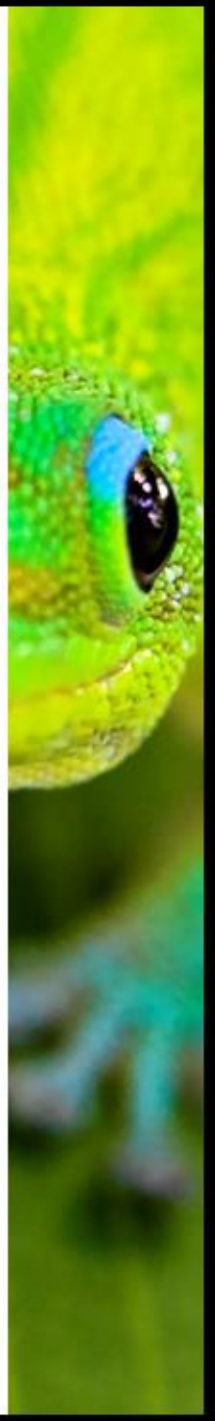
- Une **vision** ciblant la qualité des données
  - **Ne pas “réinventer la roue”** et **utiliser les standards**
  - **Chercher l’efficacité** (dans la collecte et l’assurance qualité) and **éviter la duplication d’effort**
  - **Encourager le partage** (données, informations et outils)
  - **Réfléchir à long terme**
  - **Prendre soin des utilisateurs et de leurs besoins**
  - Investir dans la **documentation** et les **métadonnées**
- Une **politique** implémentant cette vision
- Une **stratégie d’implémentation** pour cette politique (échéances précises à court, moyen et long terme)

Partage des responsabilités
Le collecteur:

- L'étiquetage est **correct**, aussi **complet** que possible et **lisible**
- Les **méthodes** de collecte sont **largement documentées**
- Les **remarques** sont **claires** et **non-ambiguës**
- ...



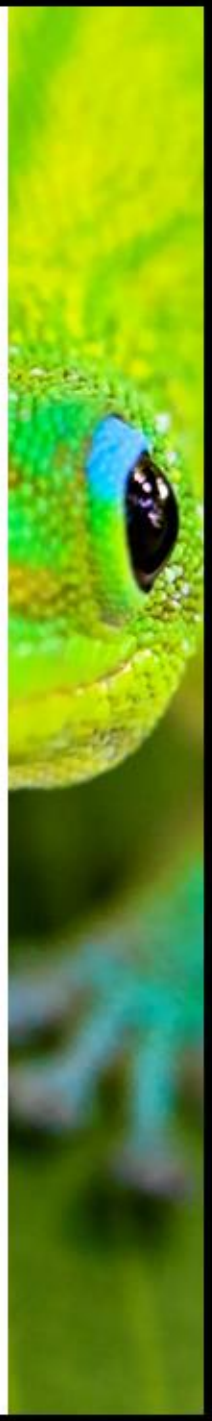
Partage des responsabilités
Le conservateur: responsabilité à long-terme



- **Qualité des retranscriptions** dans la base de données
- Des **tests de validation** sont exécutées régulièrement et documentés.
- Les données sont **sauvegardées** et **archivées**
- **Les versions précédentes** sont systématiquement **conservées**
- **Assurer le respect** (vie privées, propriété intellectuelle, sensibilités et traditions locales, ...)
- Fournir **une documentation de qualité** (incluant **les problèmes connus**)
- **Les retours utilisateurs** sont pris en compte

Responsabilité de maintenance, mais aussi la responsabilité morale d'améliorer la qualité des données (si possible) pour de futurs utilisateurs et usages.

Partage des responsabilités
L'utilisateur :



Informer les conservateurs:

- **Erreurs** et omissions dans les **données** et la **documentation**
- Définir les **priorités futures**
- ....

A l'usage:

- **Déterminer si les donnés sont adaptées à l'usage prévu** et ne pas les utiliser de façon non-adéquate.

Exactitude et précision

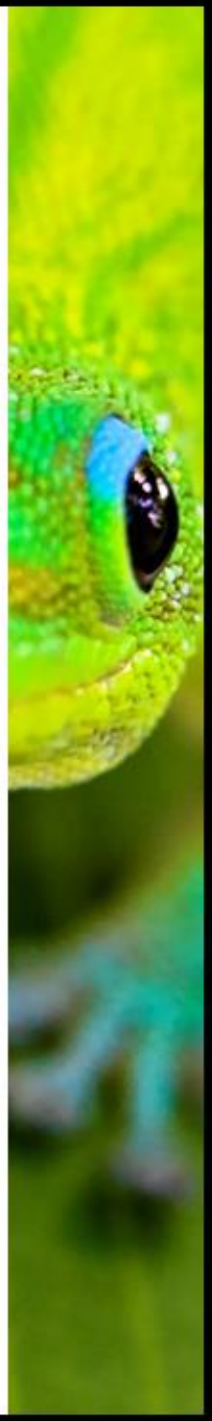
Exactitude : véracité de l'information

Précision : décrit à quel point la valeur mesurée est proche de la « vraie » valeur (statistique ou numérique)

Exactitude faible
Haute précision

Haute exactitude
Basse précision

Haute exactitude
Haute précision



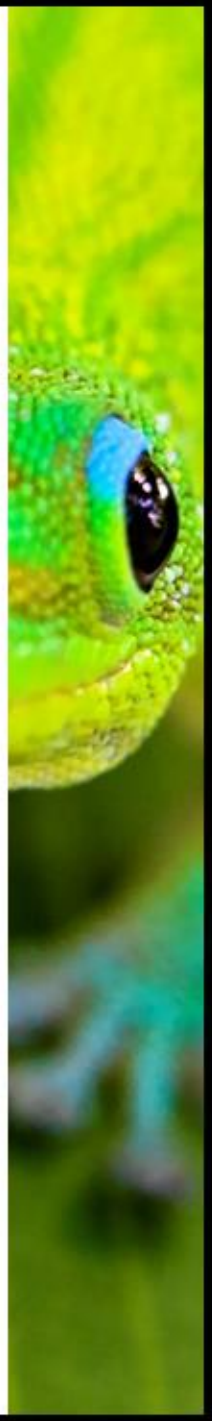
Erreur et incertitude

Erreur

- Englobe **imprécision et données inexactes**
- **Aléatoire** ou **systématique**
- **Inutile de tenter de lui échapper** (mesure, calcul, enregistre et documente)

Incertitude

- **Toujours présente** (difficulté: comprendre, décrire et enregistrer)
- Nous en dit plus sur l'**observateur** que sur les données elles-mêmes !

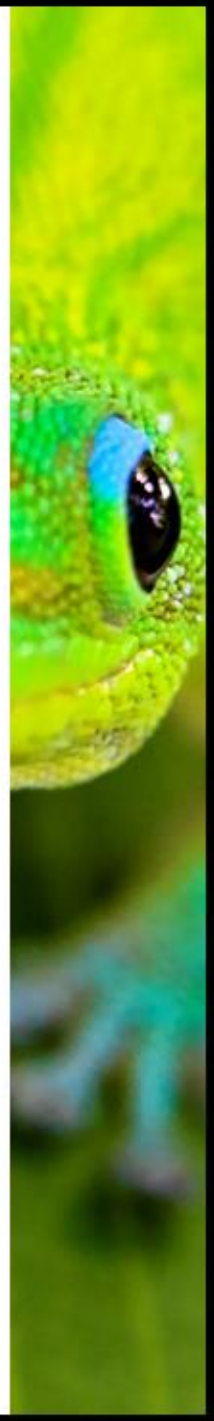


Adéquation à l'usage et métadonnées

Métadonnées = « Données sur les données »

- **Décrivent le contenu**, l'**accessibilité**, la **complétude**, ...
- A propos du **dataset**
- **Documentation de l'erreur**
- Documentation des **procédures de validation**, de **nettoyage** et de **correction** appliquées

Les métadonnées doivent être suffisament riches pour permettre l'usage des données par des tiers sans devoir se référer à la source de ces données.

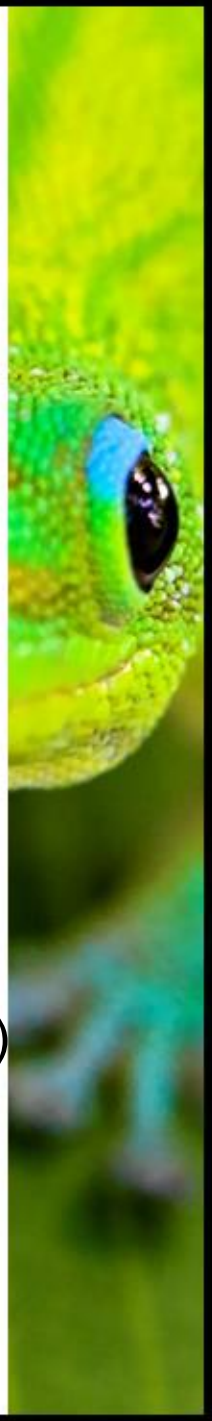


Données taxonomiques

Souvent le **nom = point d'entrée**

⇩

risque de propagation des erreurs tout au long du processus de publication des données



Erreurs possibles et solutions :

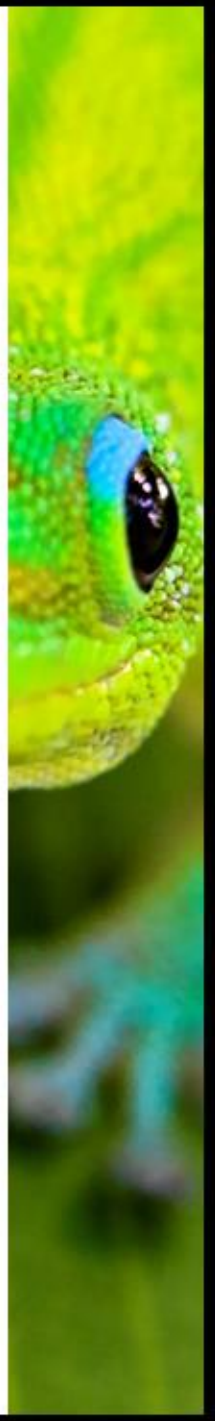
- Identification incorrectes (chercher l'aide d'un taxonomiste)
- Erreurs orthographiques (nettoyage des données)
- Mauvais format (nettoyage des données)

Les erreurs peuvent concerner noms scientifiques et noms communs, à tous les niveaux de taxonomie

Données taxonomiques

De quoi parle-t-on ?

- **Noms** (scientifique, vernaculaire, rang, hiérarchie, …)
- **Statuts** (synonymes, nom valide, …)
- **Références** (auteur, date et lieu)
- **Détermination** (par qui et quand ?)
- **Champs relatifs à la qualité** (certitude, …)



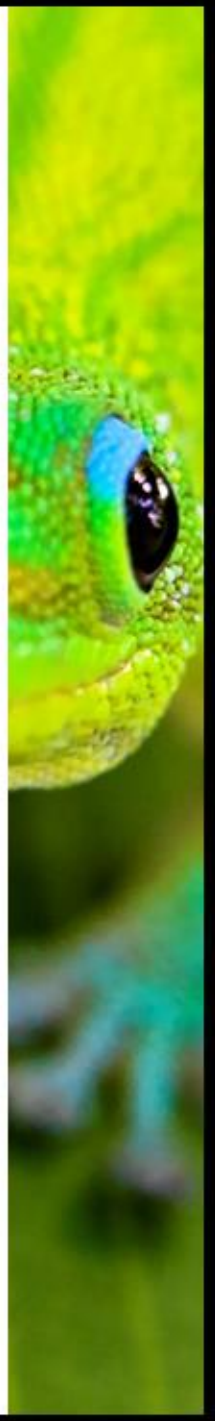
GBIF
www.gbif.org

gbif France

Données taxonomiques

Erreurs courantes

- Données manquantes (ex : sous-espèce renseignée mais pas l'espèce)
- Valeurs incorrectes (fautes de frappe, mauvaise colonne, symboles « ?? », …)
- Valeurs non-atomiques (ex : « subsp. bicostasa » dans un seul champ
- Incertitude sur un des noms de la nomenclature binomiale
- Valeurs dupliquées (synonymes, plusieurs noms valides…)
- Données inconsistantes suite à la fusion de deux bases de données utilisant différents référentiels



GBIF
www.gbif.org

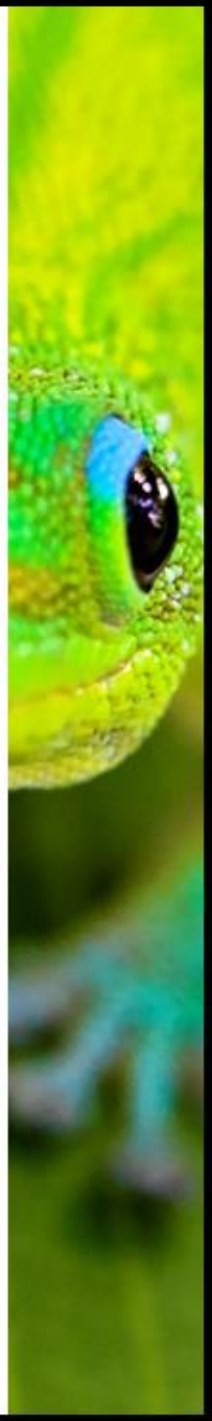
gbif France

Données spatiales

Introduction

Les données spatiales (textuelles ou géoréférencées) représentent un des aspects cruciaux pour déterminer l'adéquation à l'usage des données primaires de biodiversité:

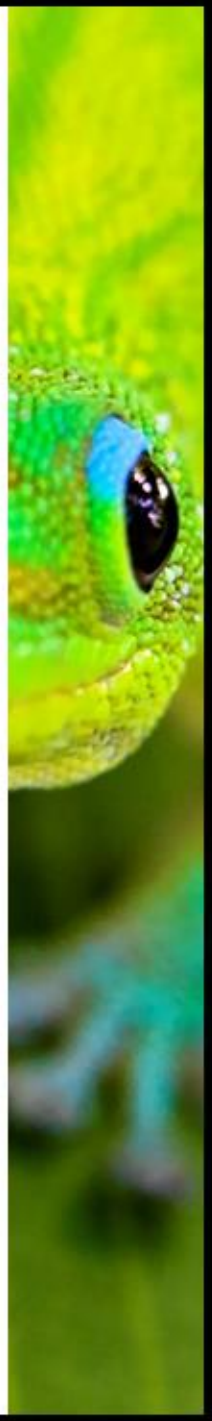
- Modélisation de la distribution des espèces
- Sélections des zones à protéger
- Gestion de l'environnement et des ressources
- ...



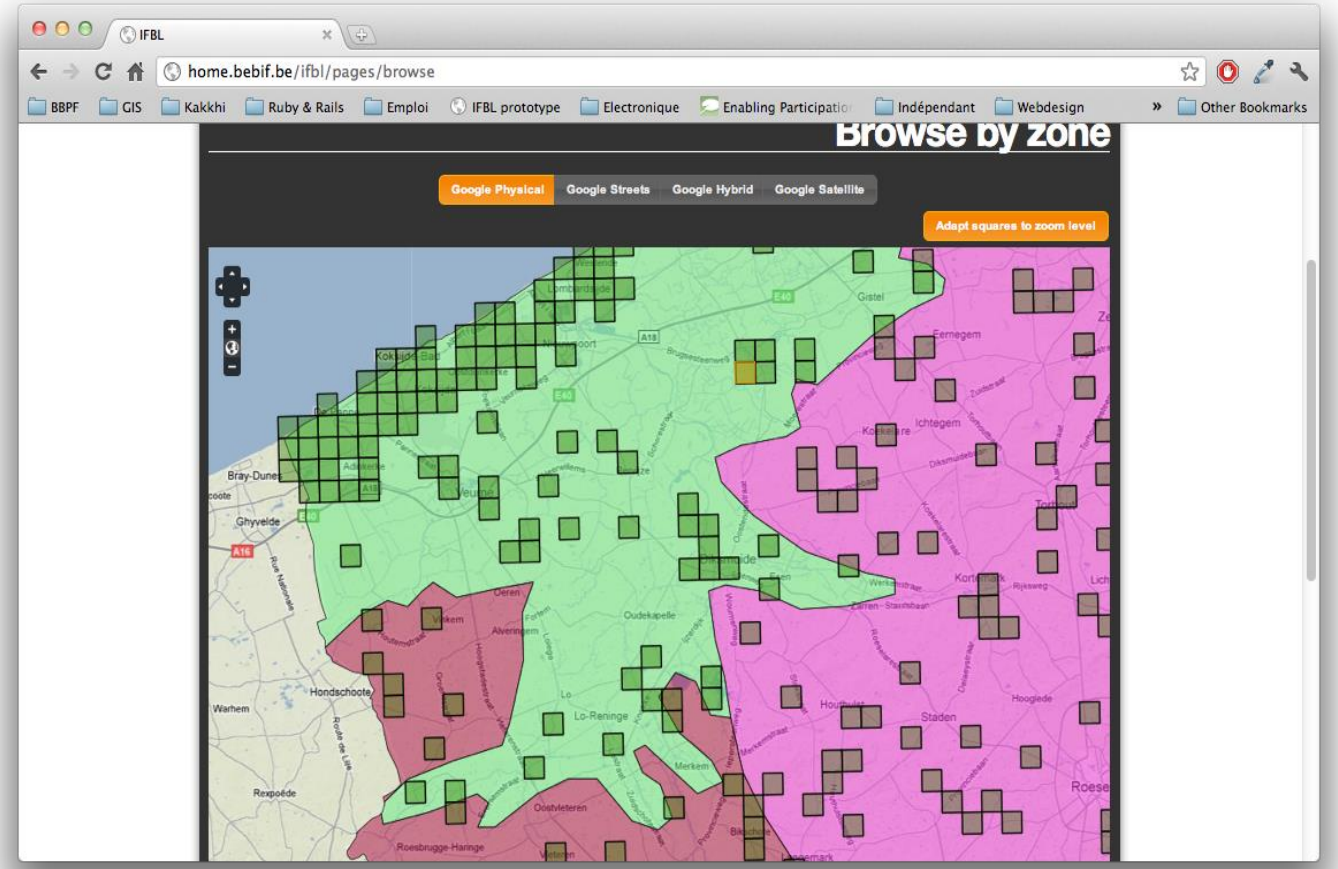
Données spatiales

De quoi s'agit-il ?

- Latitude et longitude
- Aire
- Point + rayon
- Boîte englobante (bounding box = rectangle calculé à partir des coordonnées de deux points
- Polyline
- Référence de grille



Données spatiales



Données basées sur une grille

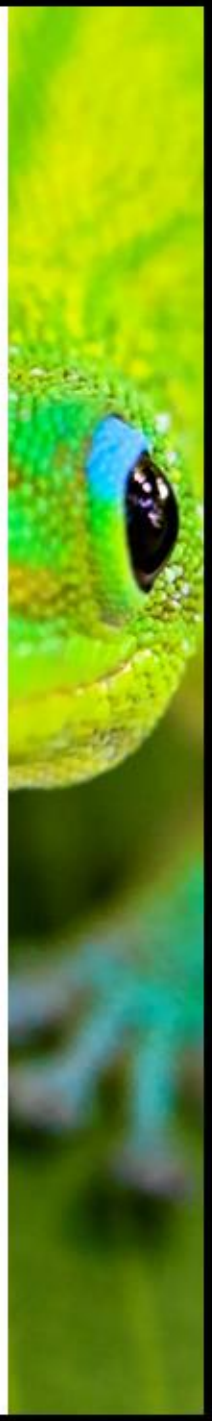
Données spatiales

Quelques définitions

- Coordonnées : un code documentant une **position sur la surface de la terre**, exprimé suivant un SRS (**spatial reference system**). En pratique; souvent latitude/longitude

- Géoréférencement : le procédé qui consiste à assigner une référence géographique à un enregistrement donné.

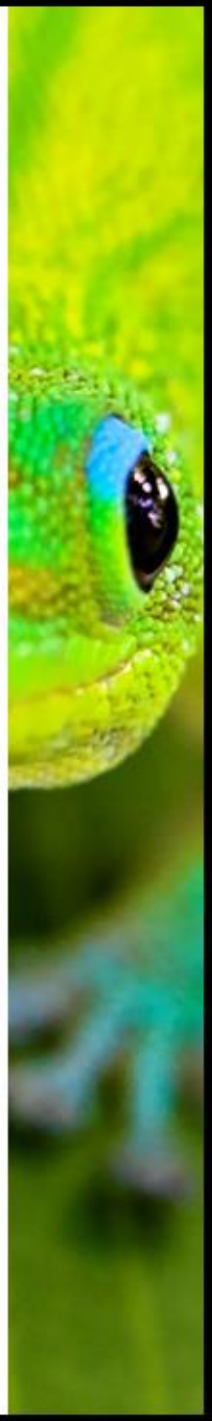
- Datum (système géodésique)



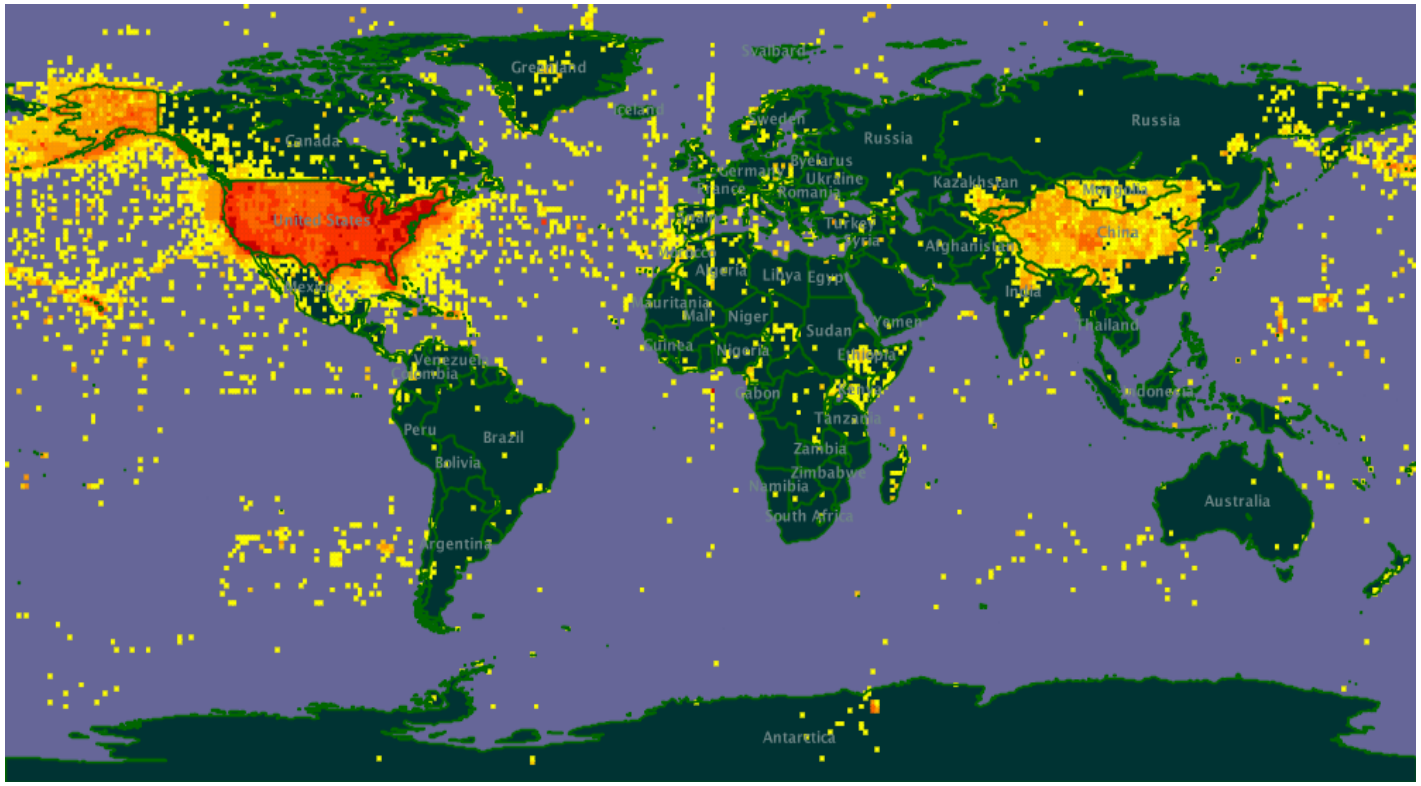
Données spatiales

Erreurs courantes

- **Inversion** des coordonnées
- Valeur(s) **zéro**
- Système géodésique/**datum inconnu**
- **SRS inadapté**
- Problèmes de **conversion**.



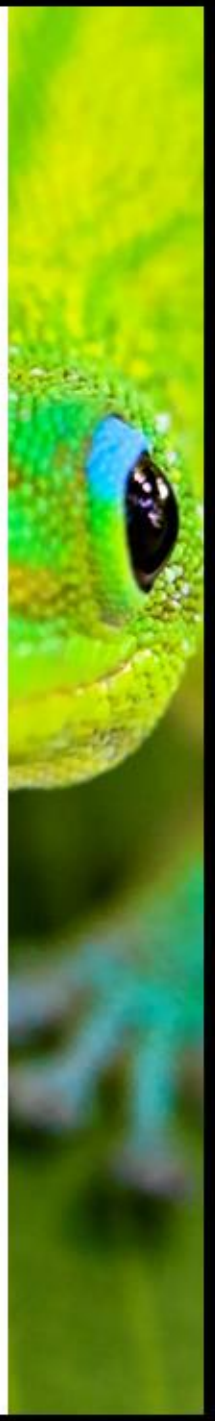
Données spatiales



Données brutes du GBIF (occurrences des USA)

Exemples d'erreurs communes :

- Coordonnées 0/0 (méridien de Greenwich et Equateur)
- Coordonnées inversées (miroir sur la Chine et léger miroir à l'ouest du Chili)



GBIF
www.gbif.org

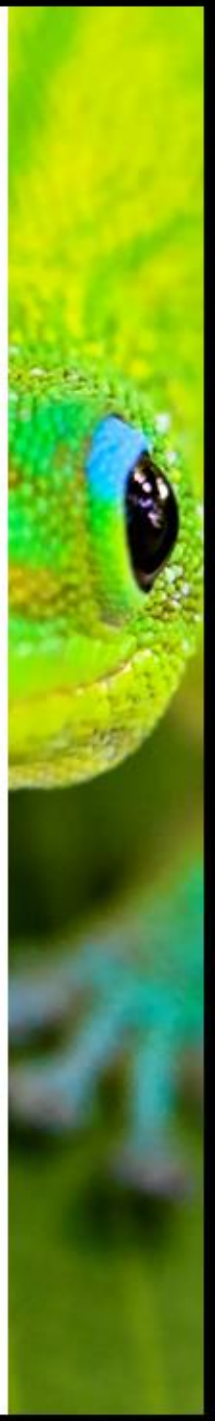
gbif France

Données de collecte et de collecteur

- **Nom du collecteur**
- **Date** de collecte
- **Informations supplémentaires**: habitat, sol, conditions météorologiques…

La pertinence dépend du type de jeu de données:

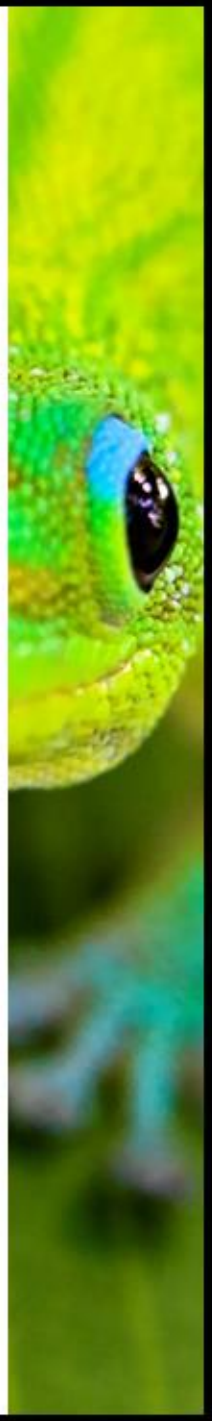
- **Collection statique (musée)** : nom et ID du collecteur, date, habitat, méthode de capture ...
- **Observations**: +durée d'observation, zone, période de la journée, activité, sexe du spécimen observé…
- **Echantillonage et inventaires exhaustifs** : +méthode, taille de la grille, fréquence, si des spécimens de référence ont été collecté (+références)



Données de collecte et de collecteur

Facteurs

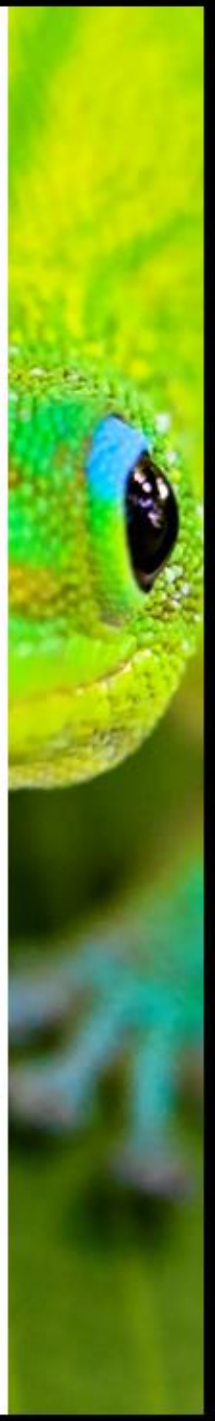
- **Exactitude**: nom du ou des collecteurs, date,…
- **Cohérence**: utilisation d'une terminologie (différente pour les sols, les habitats…)
- **Complétude** : certains champs sont très rarement renseignés (floraison, espèces associées…) ce qui peut limiter la réutilisation des données



Données descriptives

Données morphologiques, phénologiques, …

- **Qualité très variable** : données historiques impossibles à vérifier, description trop coûteuse en temps/argent, subjectivité (estimation des couleurs, de l'abondance…)
- Souvent des données s'appliquant au **niveau taxonomique** et pas au niveau du spécimen
- **Complétude** : généralement impossible à atteindre sur un même spécimen
- **Cohérence**: attributs non consistants
  - FLOWER_COLOUR = MAUVE
  - FLOWER_COLOUR= violet clair



Merci pour votre attention