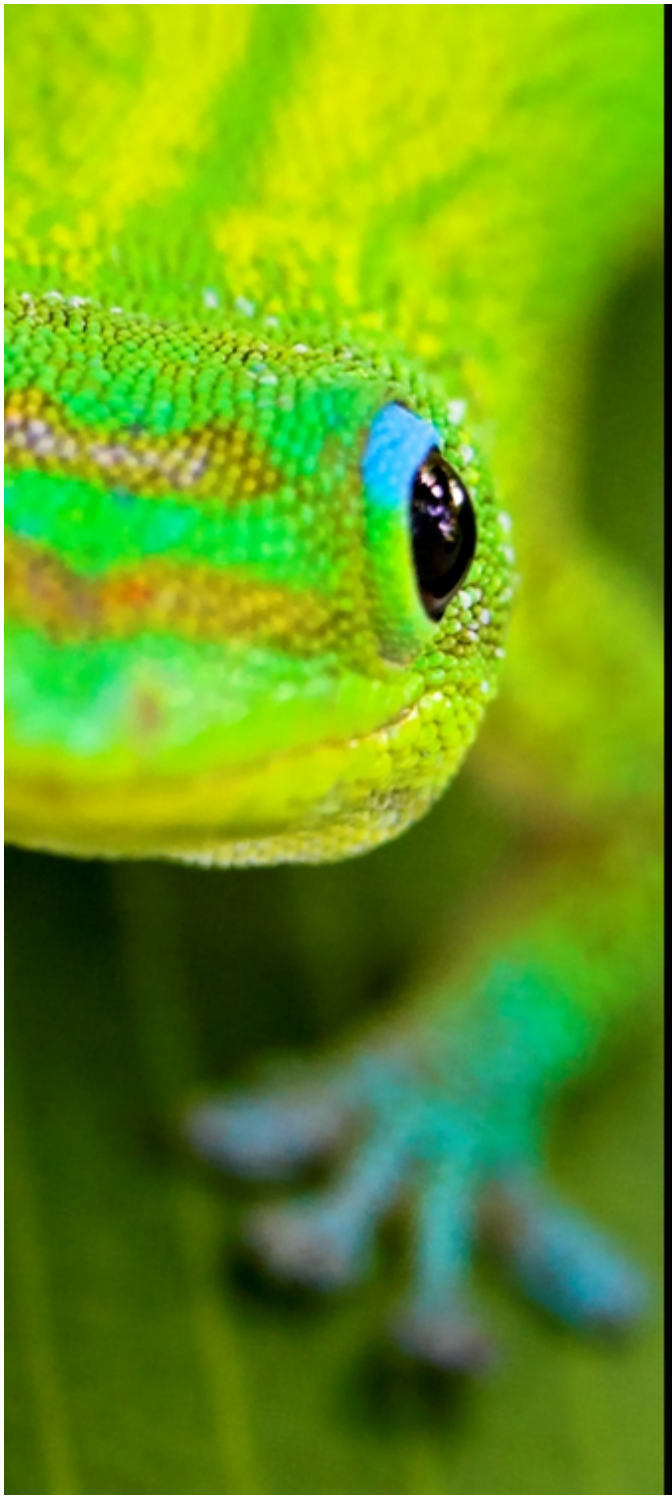


Formation qualité, publication et utilisation des données– Paris, 24-25 mars 2014

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

Sophie Pamerlon (pamerlon@gbif.fr)

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

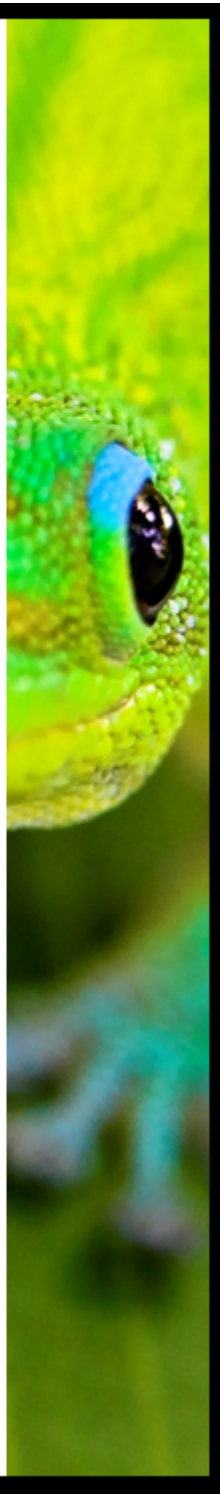
Aperçu

1. La **valeur** des données

2. **L'adéquation à l'usage**, qu'est ce que c'est ?

3. **L'Adéquation à l'usage et les données** primaires de **biodiversité** :

- Métadonnées
- Données taxonomiques
- Données spatiales
- Données sur la collecte
- Données descriptives



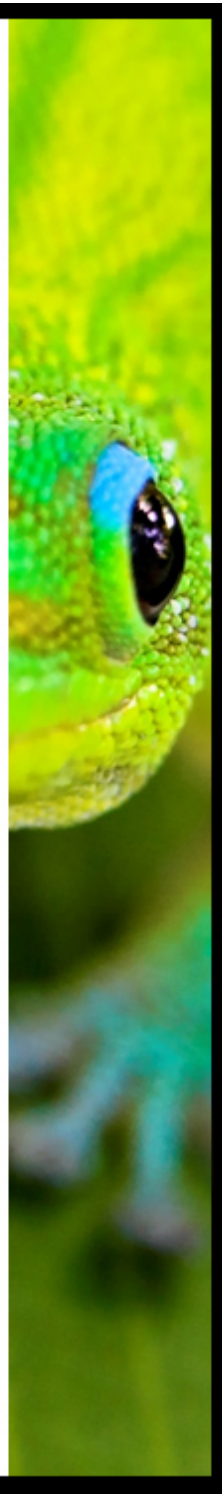
GBIF
www.gbif.org

gbif France

Pourquoi publier les données ?

21ème siècle = « siècle des données » ?

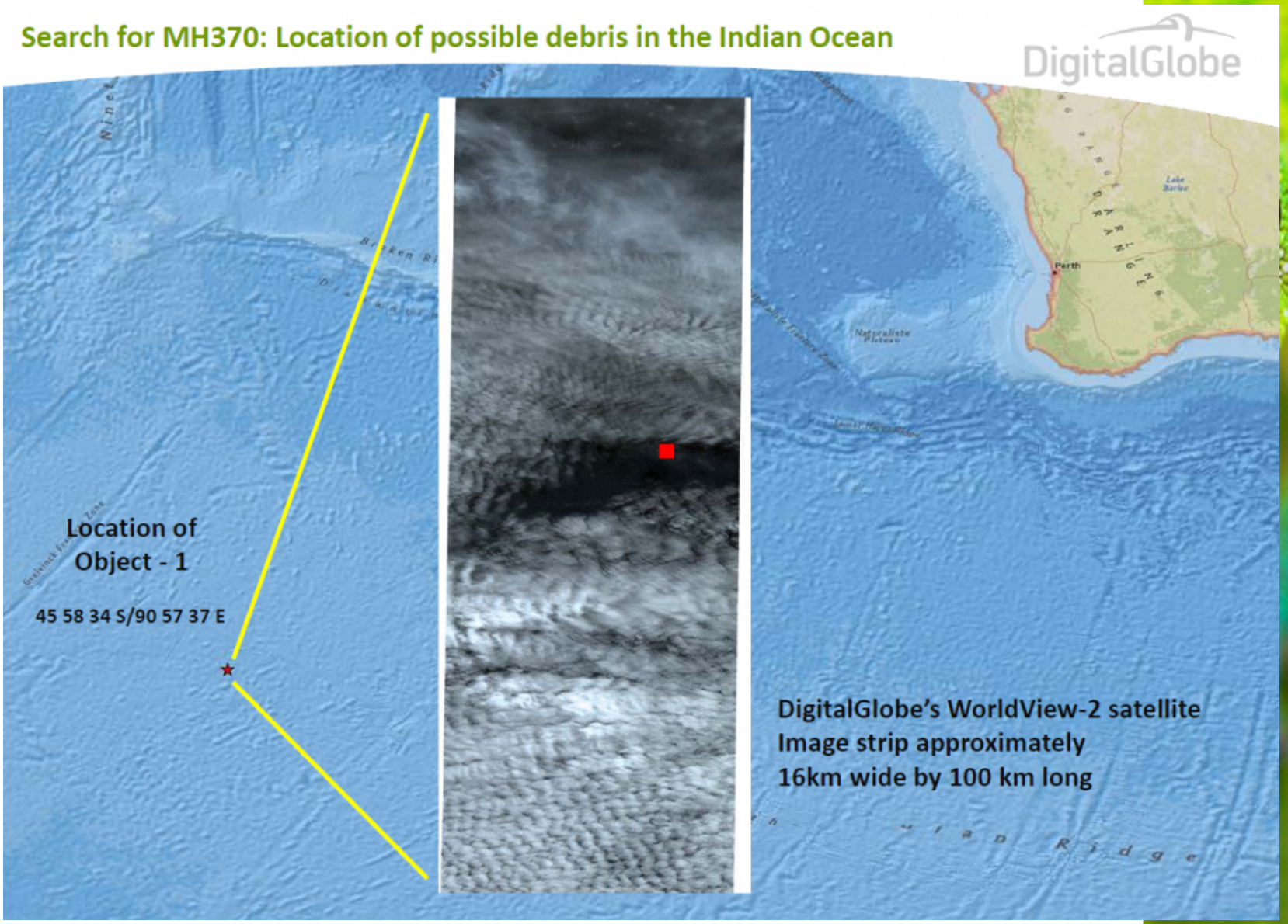
- La quantité de données augmente exponentiellement
- Le GBIF est un acteur de ce mouvement !
- Ces données **ont le potentiel** d'améliorer grandement nos connaissances et aptitudes



GBIF
www.gbif.org

gbif
France

La réponse de la communauté DigitalGlobe à la disparition du vol MH370



GBIF
www.gbif.org

gbif France

Changements climatiques et « crop wild relative »

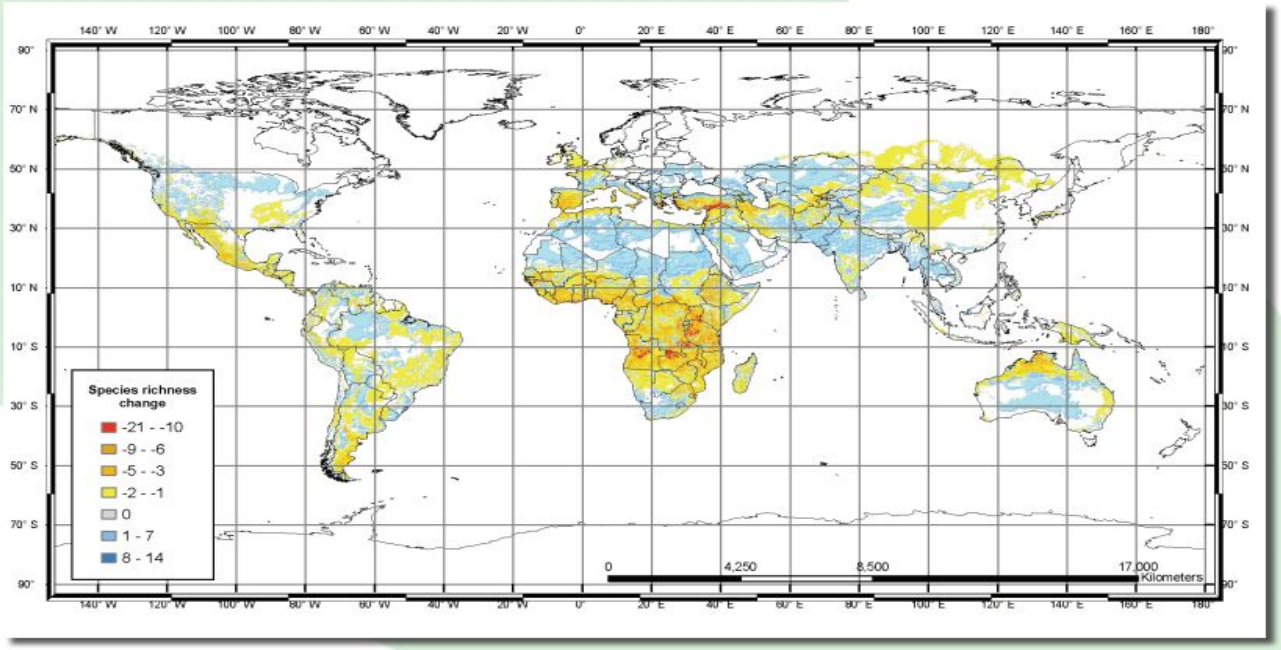
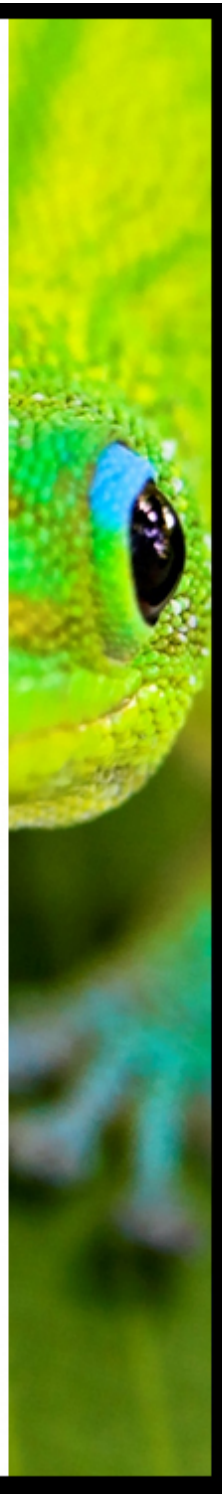


Figure 3. Predicted change in species richness to 2050

- Données du GBIF
- Crop wild relatives
- 343 espèces
- Global

- 18 modèles d'évolution climatiques
- Richesse actuelles
- Richesse future
- Prédiction du changement

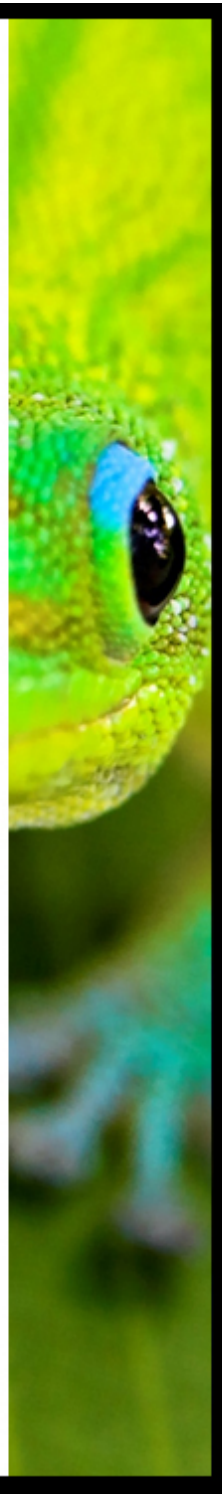


GBIF
www.gbif.org

gbif France

Des données à la compréhension…

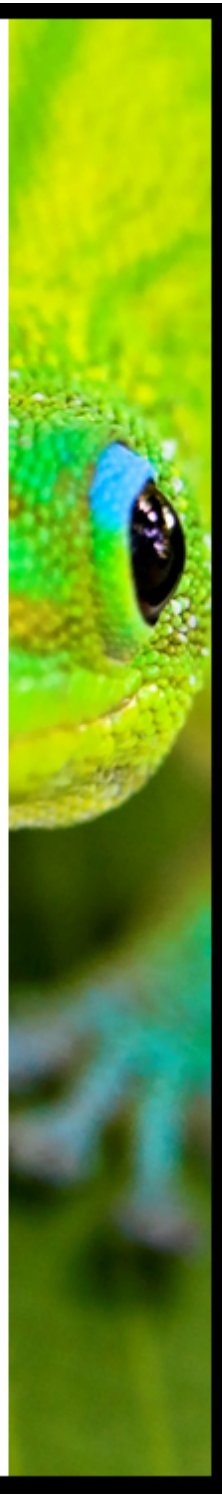
Des océans de données…



GBIF
www.gbif.org

gbif France

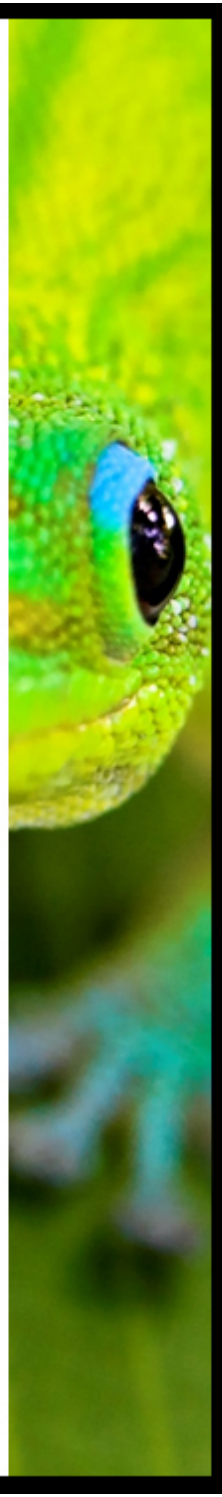
…des rivières d’informations…



GBIF
www.gbif.org

gbif France

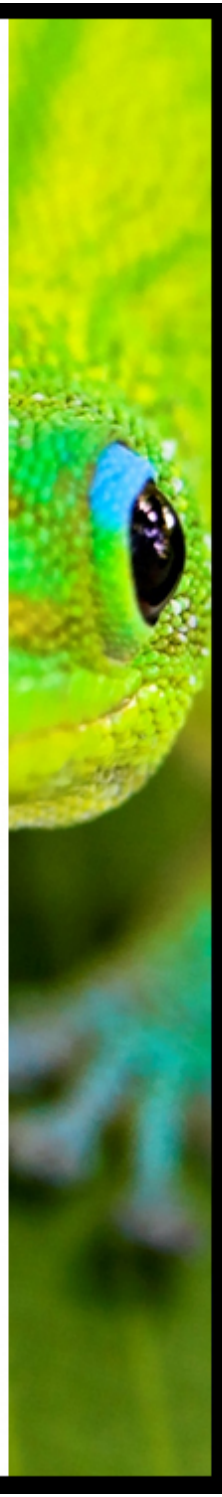
… des ruisseaux de connaissances …



GBIF
www.gbif.org

gbif
France

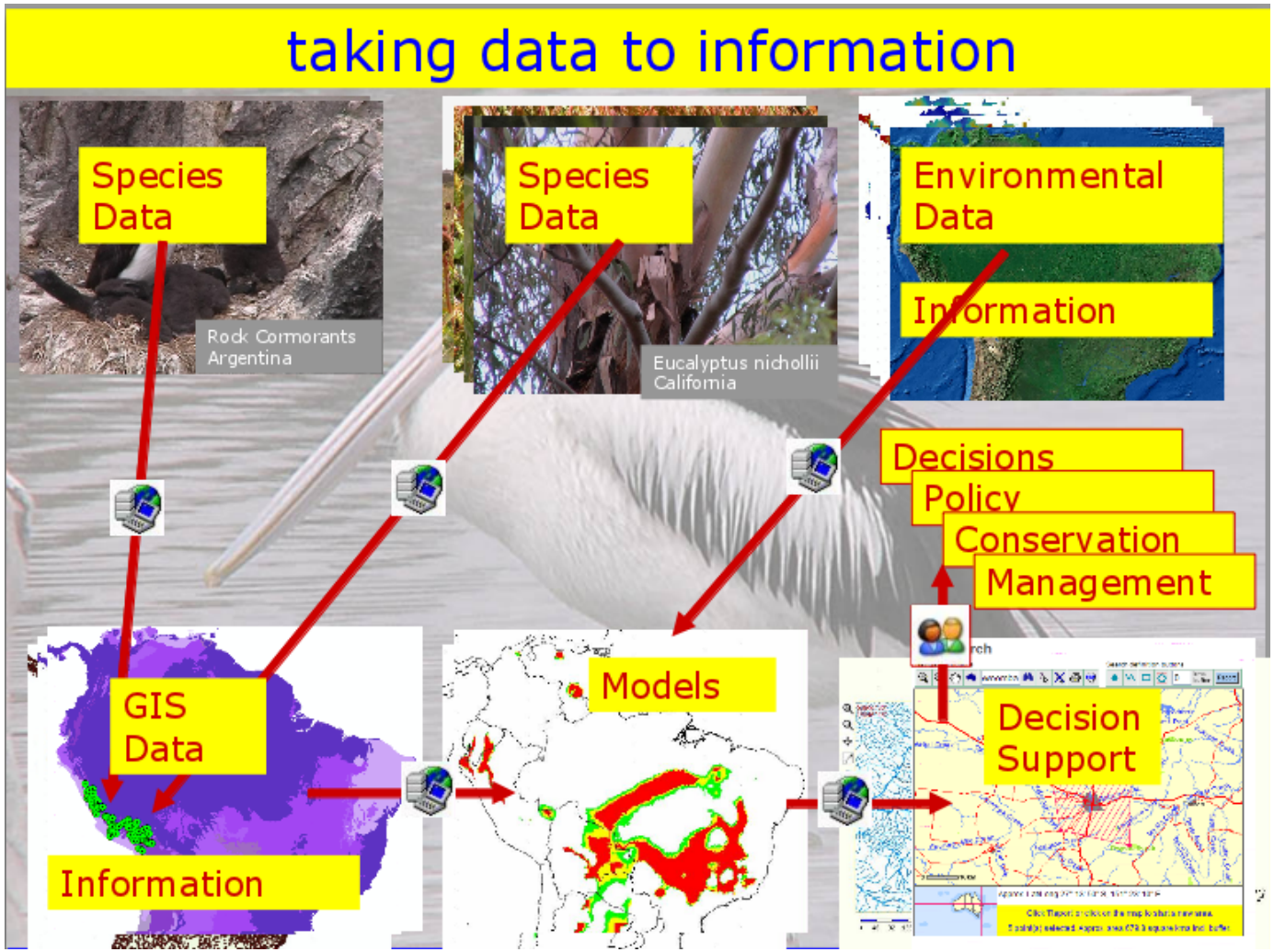
…des gouttes de compréhension



GBIF
www.gbif.org

gbif France

taking data to information



GBIF
www.gbif.org

gbif France

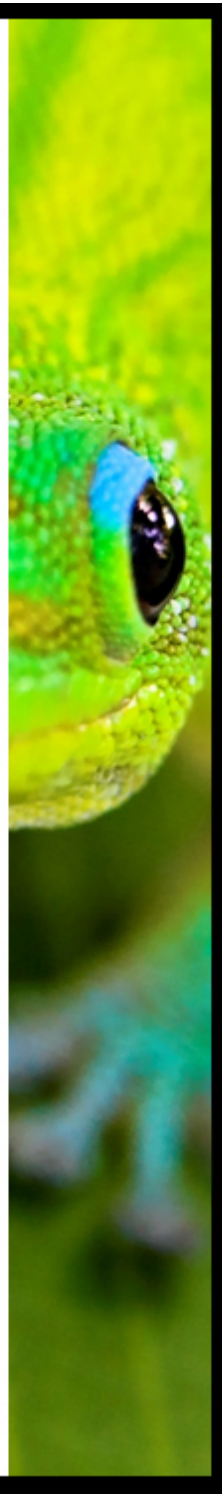
Usage des données de biodiversité

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, antivenins, éco-tourisme, histoire des sciences, chasse et pêche, rapatriation des données, photographie (et cinéma) nature, ...

D'après Chapman, 2006



GBIF
www.gbif.org

gbif France

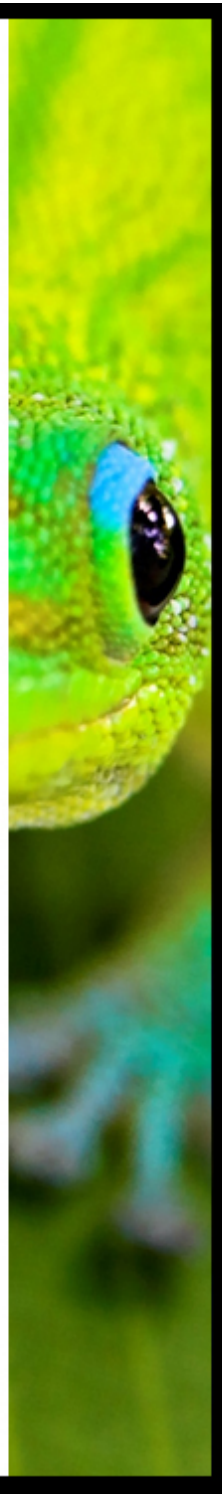
Adéquation à l'usage - définition

« 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."

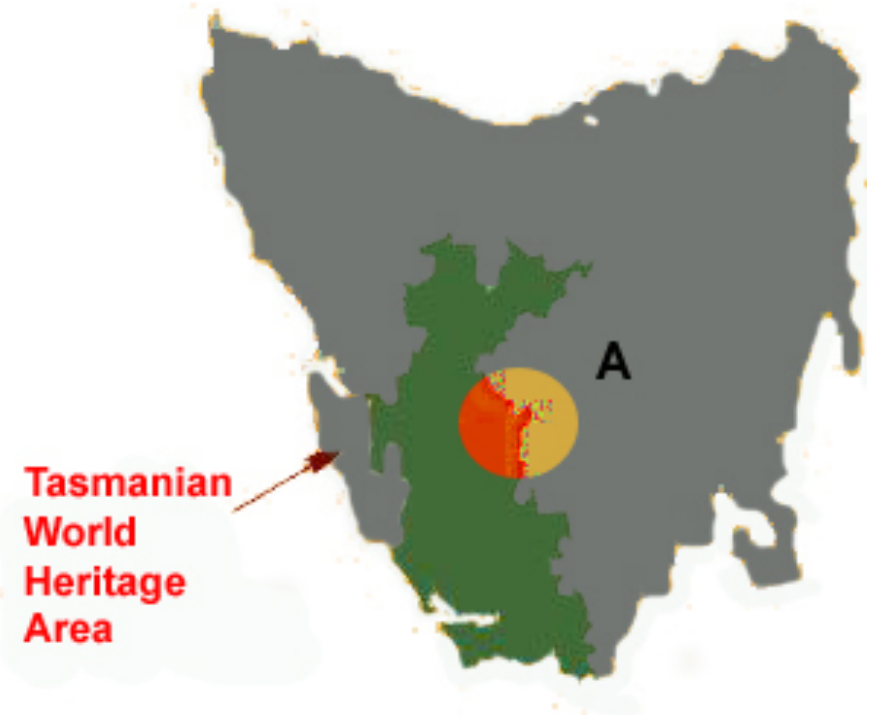
Chrisman, 1991

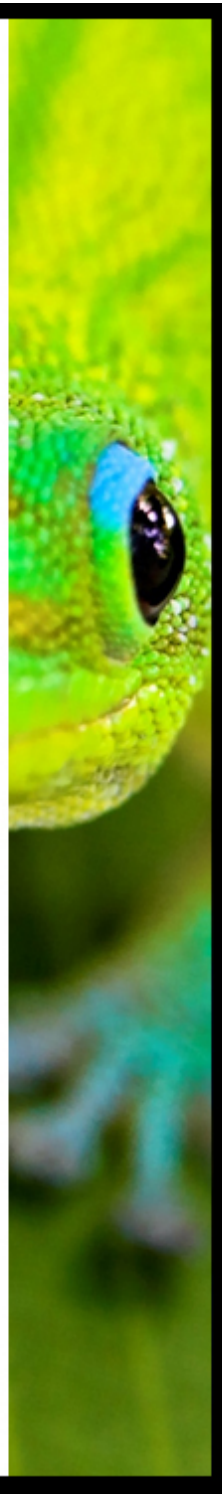


GBIF
www.gbif.org

gbif France

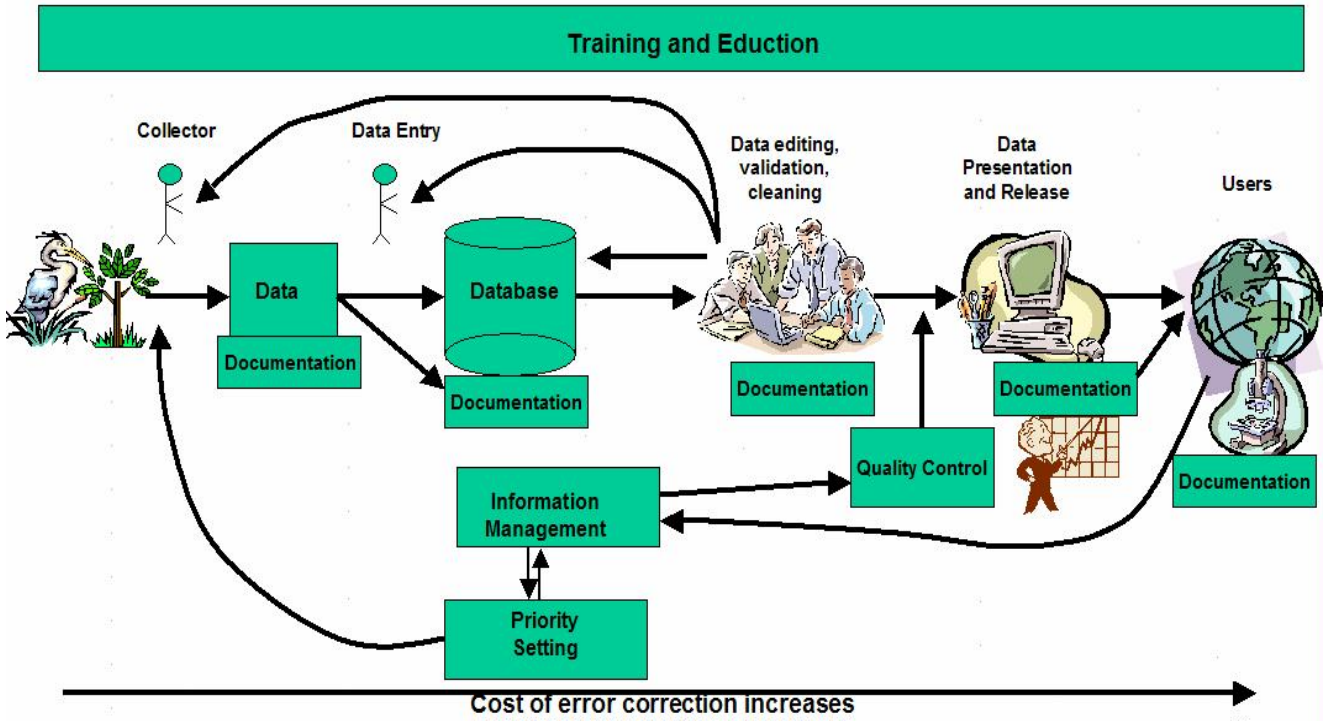
Adéquation à l'usage - exemple





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

Chaîne des données et qualité

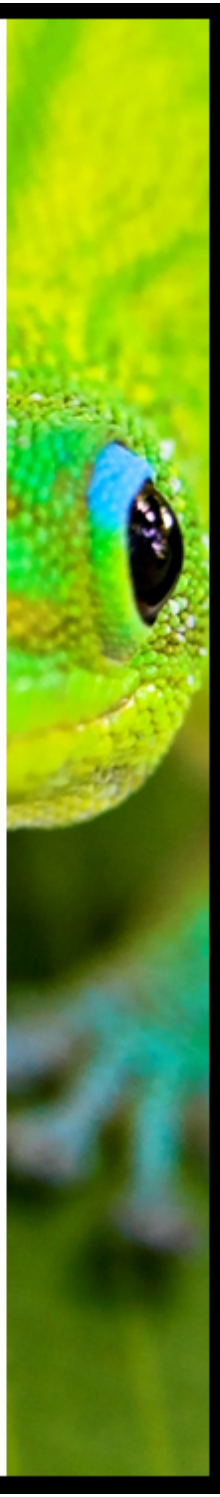


La perte de qualité survient à chaque étape.

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

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



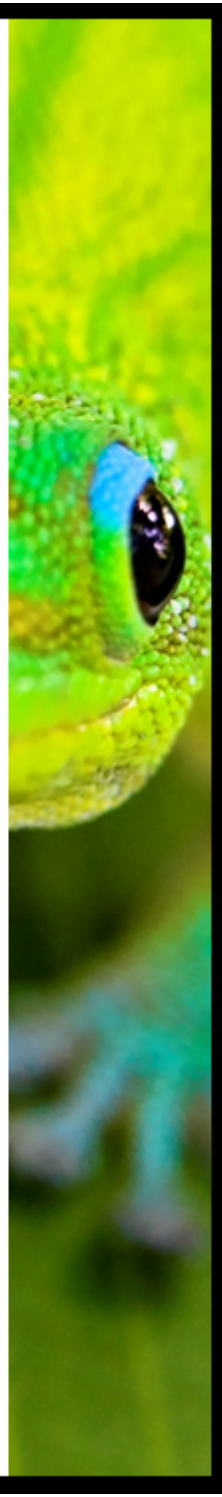
GBIF
www.gbif.org

gbif France

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**
- ...



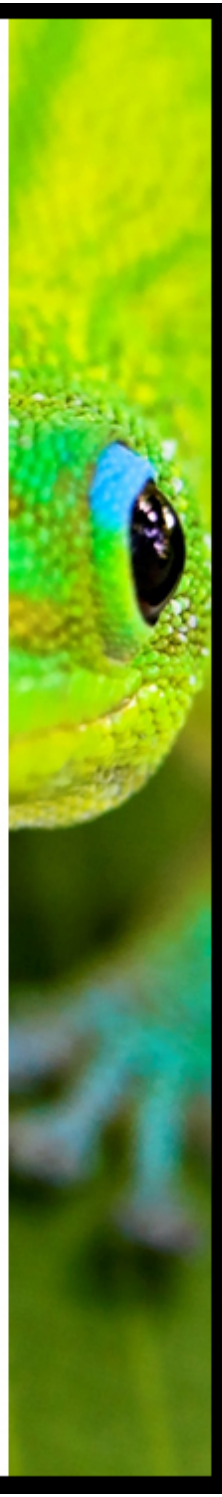
GBIF
www.gbif.org

gbif France

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, sensibilité et tradition des peuples indigènes, ...)
- 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

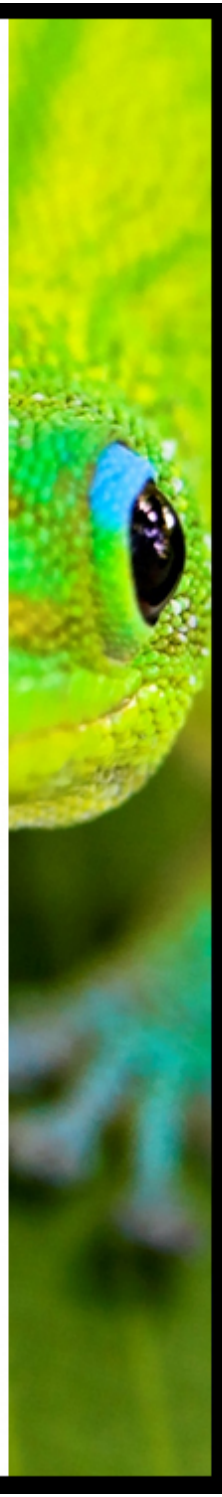
L'utilisateur:

<u>Informer les conservateurs:</u>

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

<u>A l'usage:</u>

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



GBIF
www.gbif.org

gbif France

Exactitude et précision

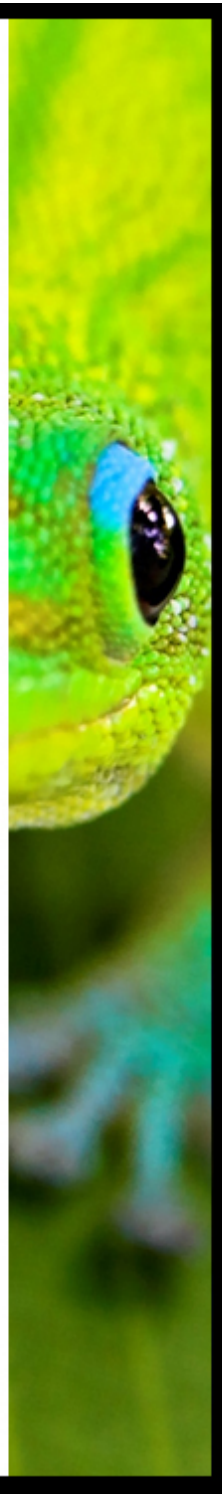
Exactitude = véracité de l'information

Précision
- Statistique
- Numérique

Exactitude faible
Haute précision

Haute exactitude
Basse précision

Haute exactitude
Haute précision



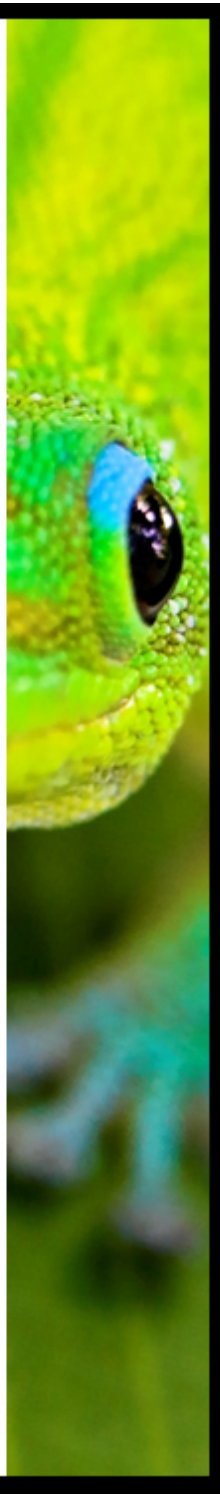
Erreur et incertitude

- Erreur : englobe **imprécision et données inexactes**
- **Aléatoire** ou **systématique**
- **Inutile de tenter de lui échapper** (mesure, calcule, entregistre et documente)

Incertitude

• Toujours présente (difficulté: comprendre, décrire et enregistrer)

• Nous en dit plus sur l'observateur que sur les données elles-mêmes !



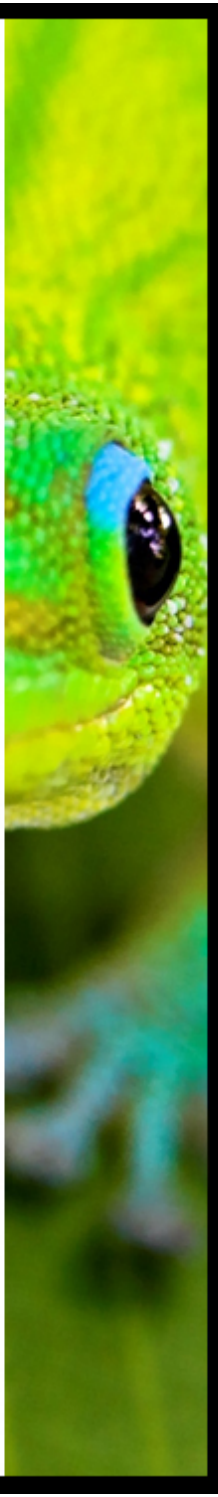
GBIF
www.gbif.org

gbif France

Adéquation à l'usage et métadonnées

"**Données sur les données**"

- **contenu**, **accessibilité**, **complétude**, ...
- A propos du **dataset** ou de l'**enregistrement**
- **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.

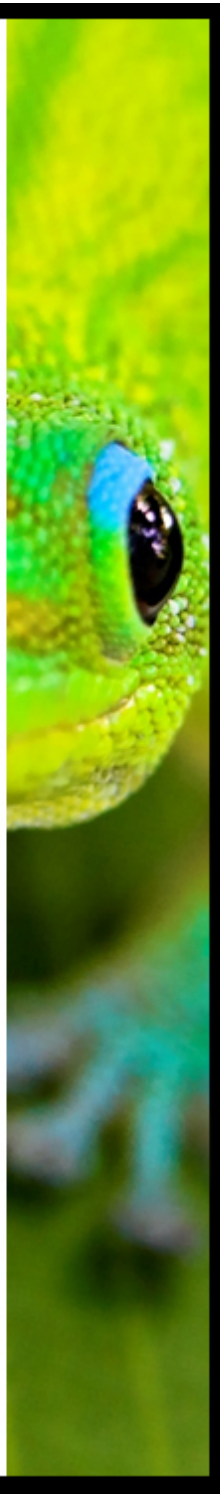
Données taxonomiques

Souvent: **nom = point d'entrée**

risque de propagation des erreurs

Erreurs possibles:

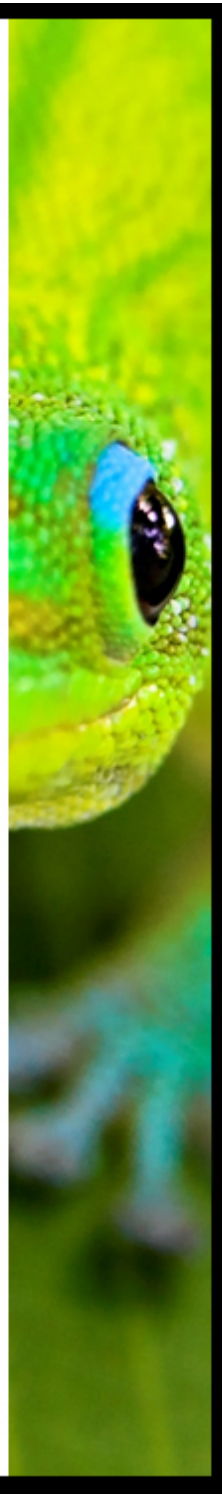
- Identification incorrectes
- Erreurs orthographiques
- Mauvais format



Données taxonomiques

De quoi parle-t-on ?

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



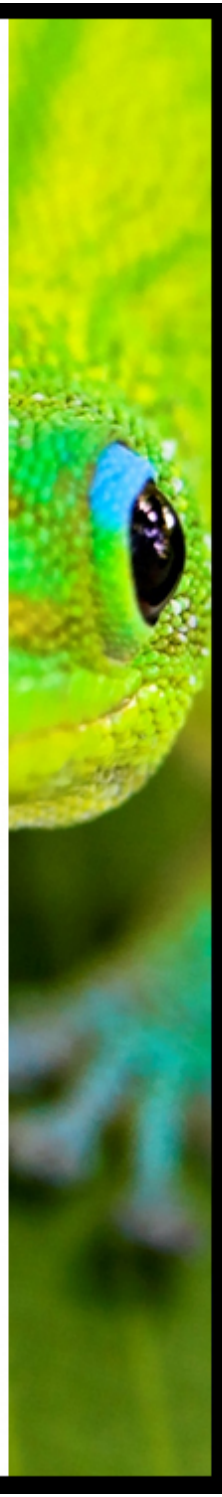
GBIF
www.gbif.org

gbif France

Données taxonomiques

Erreurs courantes

- **Données manquantes**
- **Valeurs incorrectes**
- **Valeurs non-atomiques**
- **“Domain schizophrenia”**
- **Valeurs dupliquées**
- **Données inconsistantes**

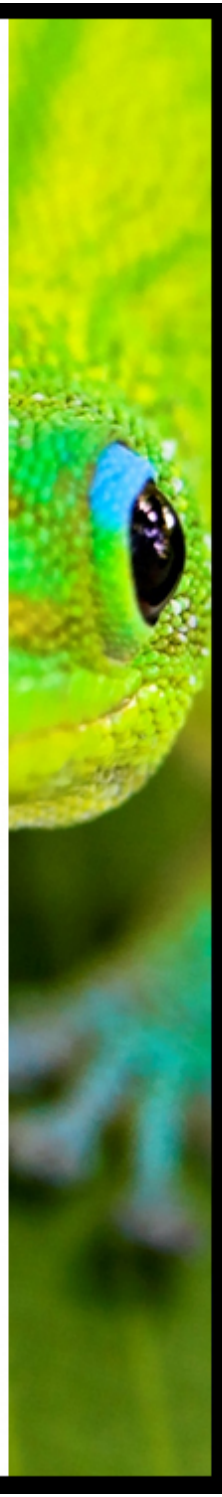


Données spatiales

Introduction

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
- ...



GBIF
www.gbif.org

gbif France

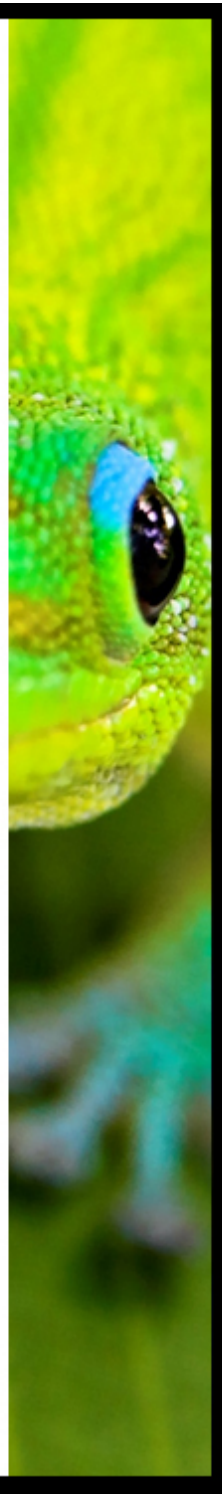
Données spatiales

De quoi s'agit-il ?

Latitude et longitude ?

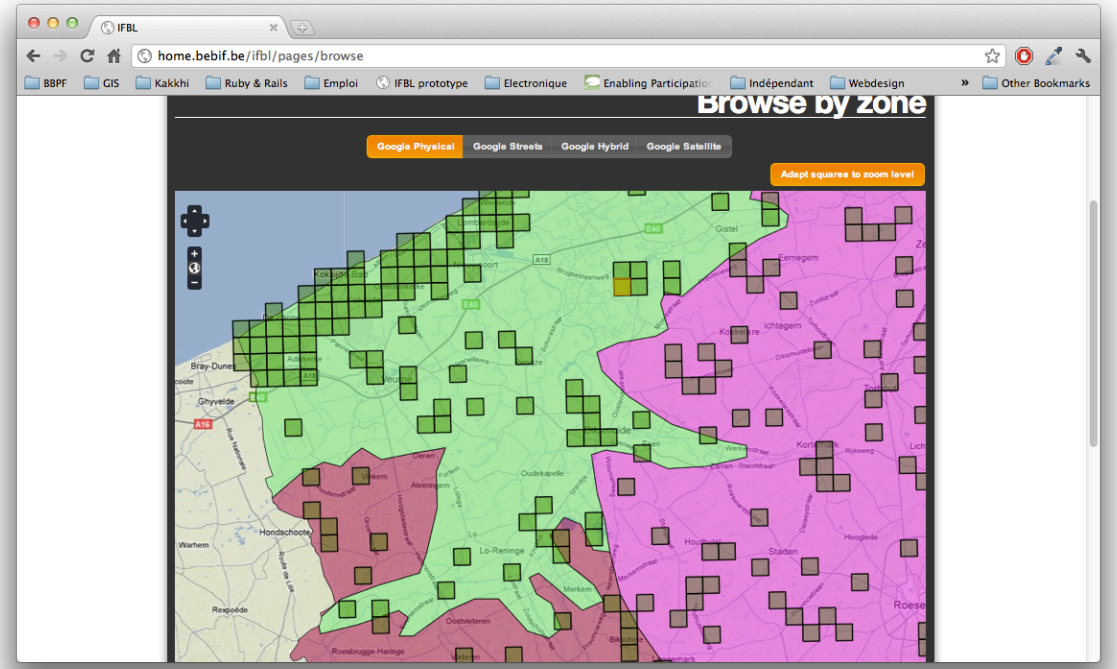
Aire !

- Point + rayon
- Boîte englobante (bounding box)
- Polyline
- Référence de grille

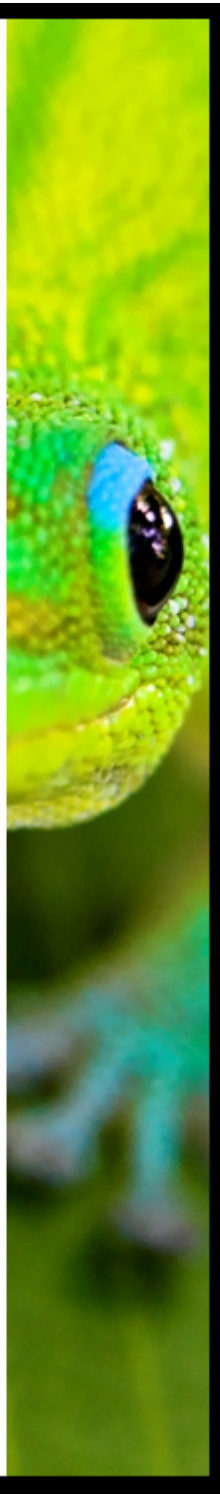


GBIF
www.gbif.org

gbif France



Données basées sur une grille (cheklists)



GBIF
www.gbif.org

gbif
France

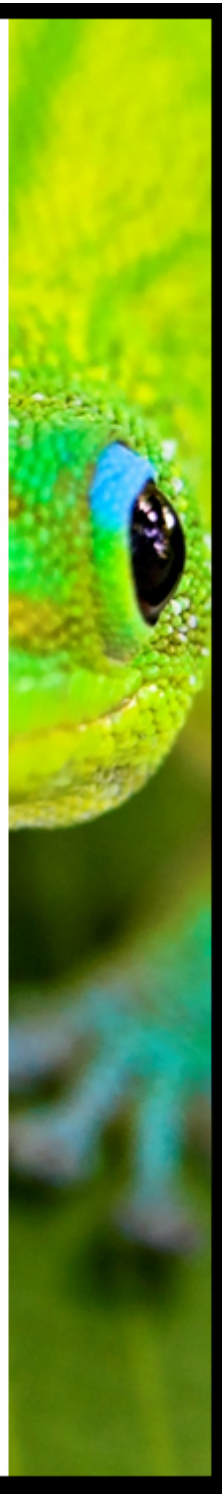
Données spatiales

Quelques définitions

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

- Géoréférencer / géocoder : le procédé qui consiste à assigner une référence géographique à un enregistrement donné.

- Datum (système géodésique)



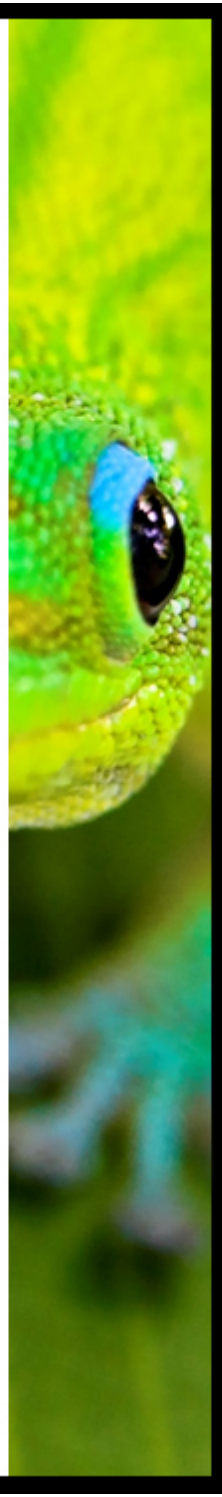
GBIF
www.gbif.org

gbif France

Données spatiales

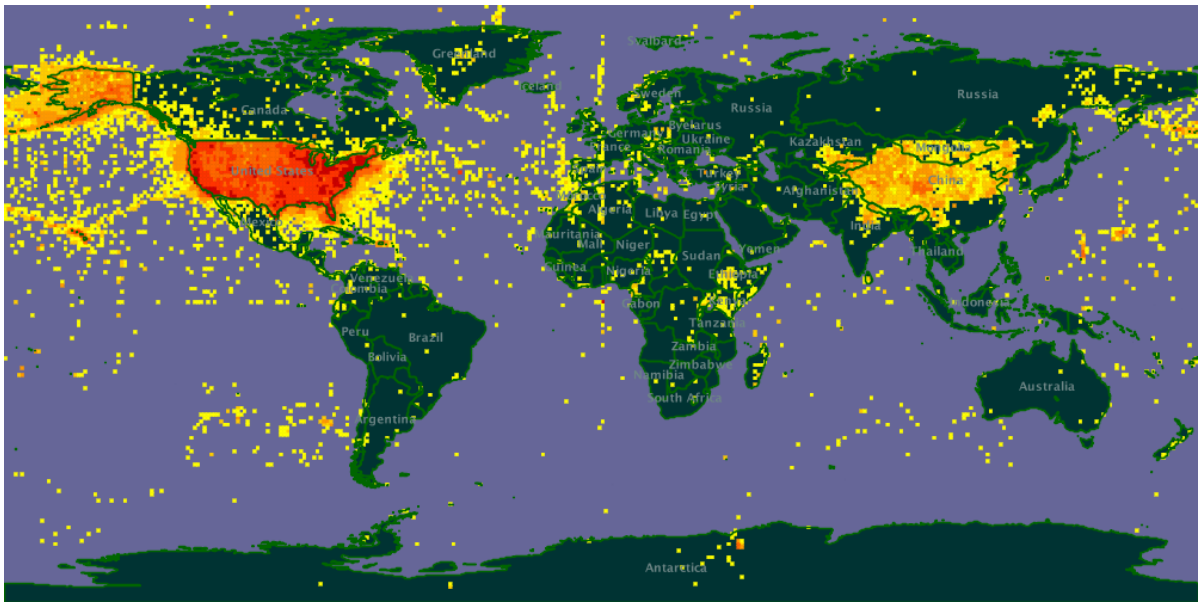
Erreurs courantes

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

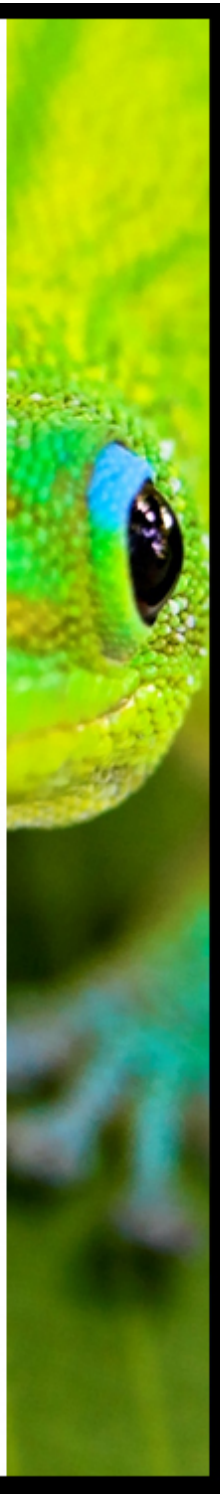


GBIF
www.gbif.org

gbif France



Données brutes du GBIF (occurrences des USA)



GBIF
www.gbif.org

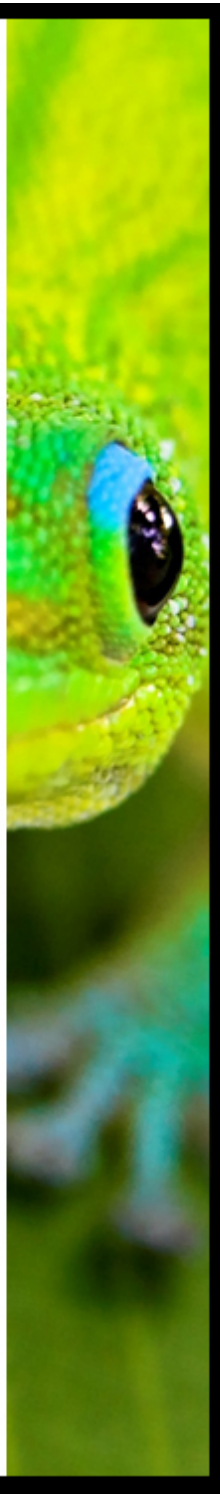
gbif
France

Données de collecte et de collecteur

- **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…
- **Sondage exhaustif**: +méthode, taille de la grille, fréquence, si des spécimens de référence ont été collecté (+références)



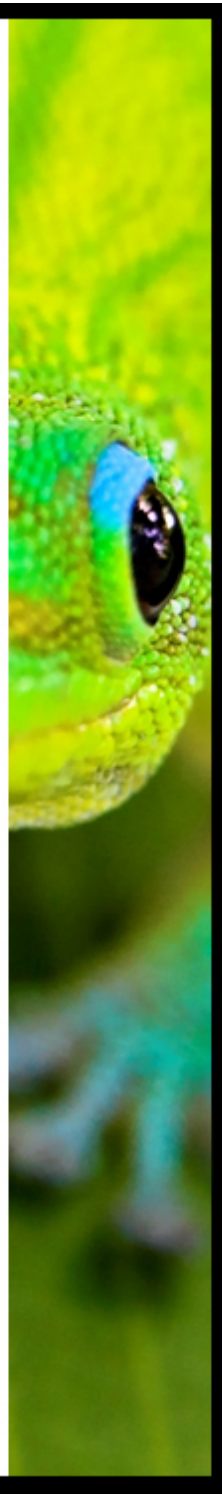
GBIF
www.gbif.org

gbif France

Données de collecte et de collecteur

Facteurs

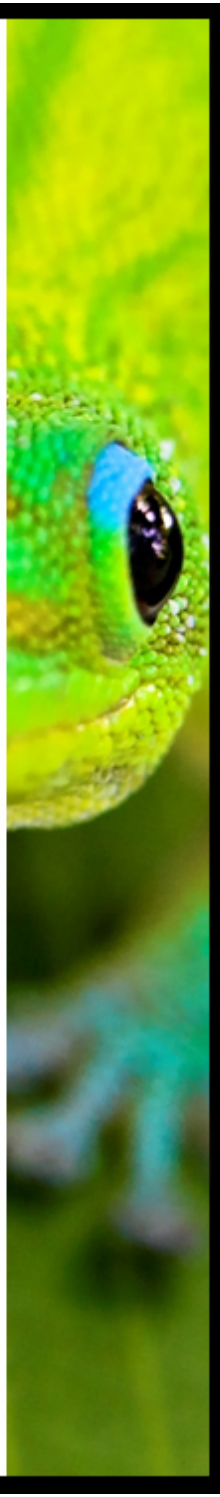
- **Exactitude**: nom de collecteurs, date,…
- **Cohérence**: utilisation d'une terminologie
- **Complétude**



Données descriptives

Morphologiques, phénologiques, …

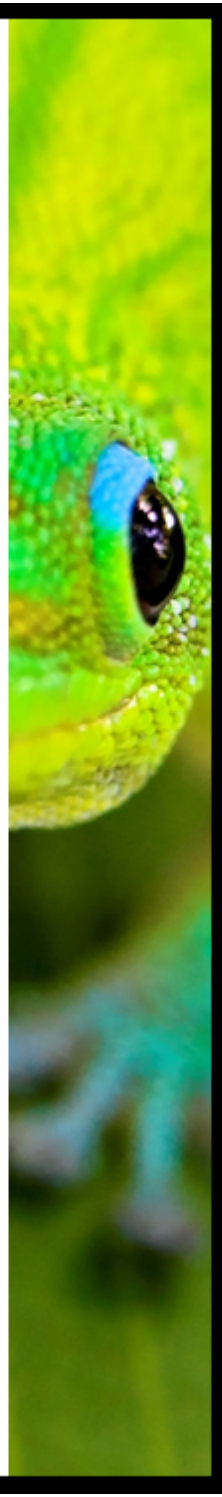
- **Qualité très variable**
- Souvent de données s'appliquant au **niveau taxonomique** et pas au niveau du spécimen
- **Complétude**: généralement impossible à atteindre sur un même spécimen
- **Cohérence**: attributs non consistants
  - FLOWER_COLOUR = MAUVE
  - FLOWER_COLOUR= violet clair



GBIF
www.gbif.org

gbif France

Questions



GBIF
www.gbif.org

gbif France

Merci

Références

Basé principalement sur les différentes présentations et publications d'Arthur Chapman

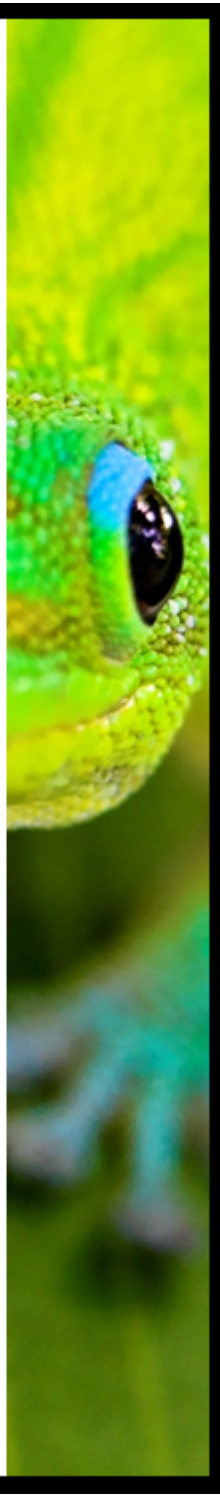
Image « point d'interrogation » par Milos Milosevic (http://www.flickr.com/photos/21496790@N06/)

Crop Wild Relatives: Andy Jarvis(1), Samy Gaiji (2), Julian Ramirez (1) and Emmanuel Zapata (1)
1. The International Center for Tropical Agriculture (CIAT)
2. The Global Biodiversity Information Facility Secretariat (GBIF)

Accuracy VS precision slide: http://www.mathsisfun.com/accuracy-precision.html

Beach picture by Lali Masrieta :www.visualpanic.net
River: **Johan J.Ingles-Le Nobel**
Stream: **bterrycompton**

Chapman, A.D. and J. Wieczorek (eds). 2006. Principes de la bonne pratique sur le géoréférencement, version 1.0. Trad. Chenin, C. Copenhague: Global Biodiversity Information Facility, 95 pp. Disponible en ligne sur http://links.gbif.org/gbif_georeferencement_manual_fr_v1.pdf



GBIF
www.gbif.org

gbif France