Formation GBIF France dans le cadre d'Ecoscope – Valoriser ses données d'observation sur la biodiversité : qualité, standards et publication
Paris, 15-16 septembre 2015

Méthodes et outils pour améliorer la qualité des données de biodiversité

GBIF France (gbif@gbif.fr)

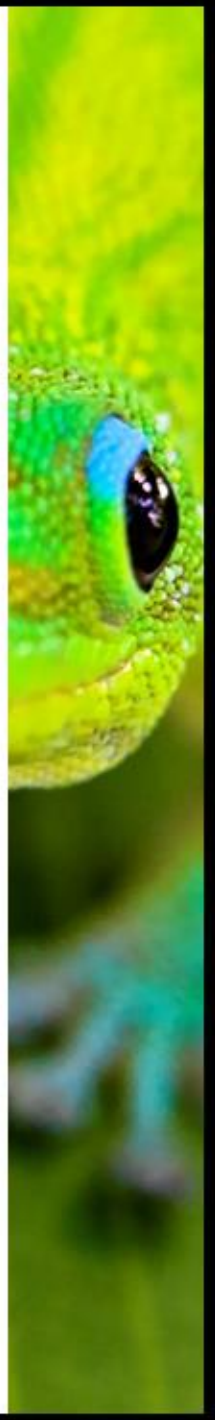
Basé sur la présentation de Nicolas Noé – niconoe@ulb.ac.be
pour GB18 training sessions – Buenos Aires, Argentine (sept 2011

Aperçu

- Guide des bonnes pratiques
  - Données taxonomiques
  - Données spatiales / géographiques
- Données sensibles
- Spécificités GBIF

Bonnes pratiques

Pour les données taxonomiques



Données taxonomiques

Certitude d'identification

Conception de la base de données:

- **Flag** de vérification, **nom** et **date**
- **Attention aux termes** "aff.", "cf.", "s.lat", ...
- Si pas identifié par expertise taxonomique, enregistrer l'information:
  - Clés taxonomiques
  - Analyses ADN
  - Révision d'un groupe taxonomique
  - ...

Données taxonomiques

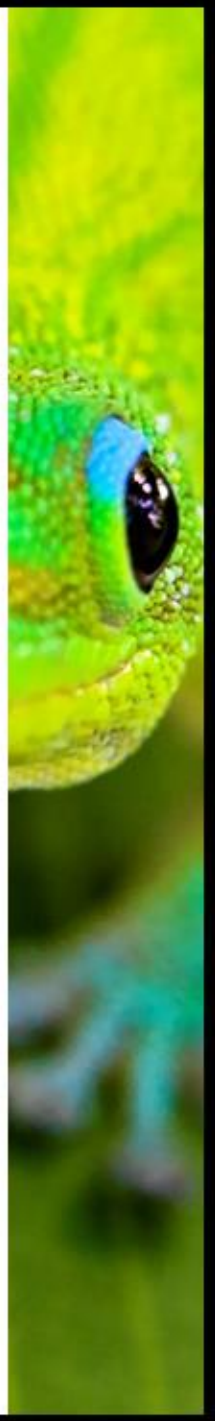
Certitude d'identification

Saisie des données:

- Utilisation de checklists
- Utilisation de fichiers d'autorité

Détection d'erreurs:

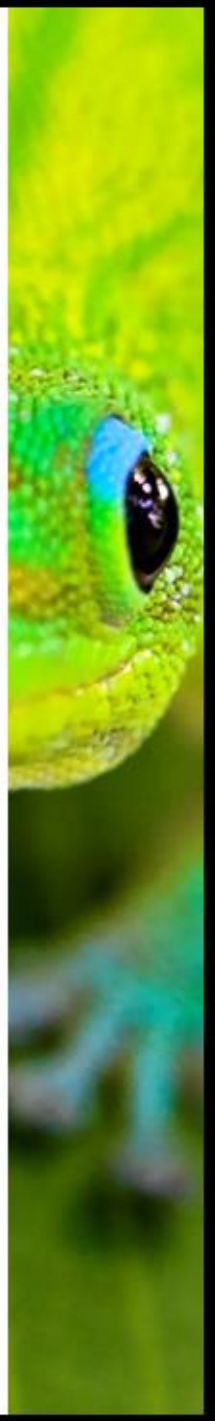
- Nécessite généralement un expert
- Les valeurs géographiques ou environnementales extrêmes (outliers) peuvent aider à détecter les déterminations aberrantes



Données taxonomiques

Erreurs orthographiques – nom scientifique

- Conception de la base de données
  - Standardiser au maximum
- Fichiers d'autorité
  - Globaux, régionaux ou par groupe
- Duplicatas
  - Interface dédiée pour la détection (+flag)

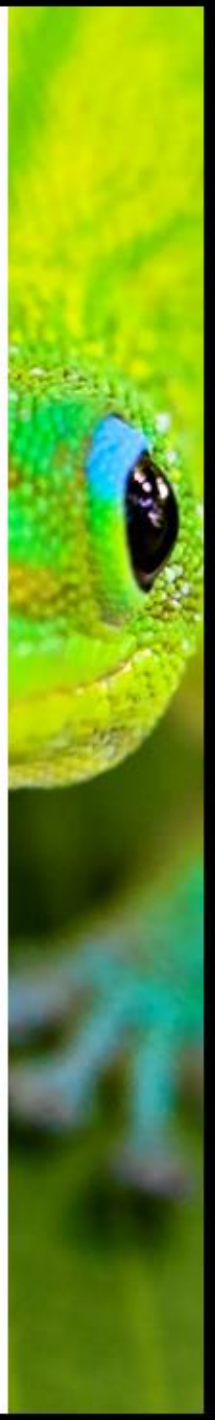


GBIF
www.gbif.org

gbif France

Données taxonomiques

Erreurs orthographiques – rang infra-spécifique



Standardiser !

Genus	Espèce	Rang_infra	Val_infra
Stipiturus	malachurus	Subsp.	parimeda

Toujours séparer rang (sp, subsp.,) et valeur (« parimeda ») pour

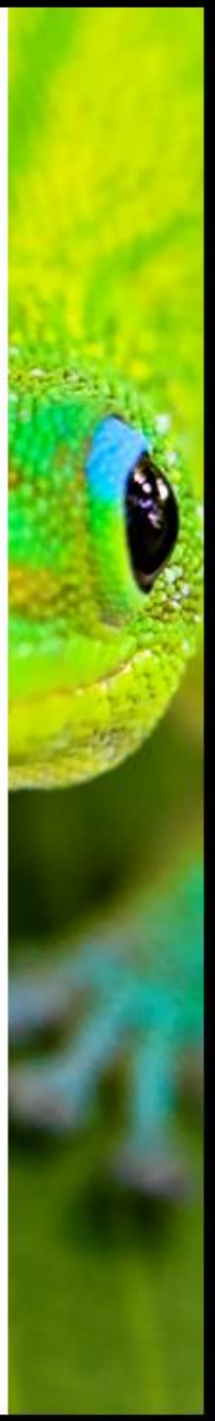
- Éviter les ambiguïtés
- Faciliter les vérifications

Données taxonomiques

Rang infra-spécifique- saisie des données

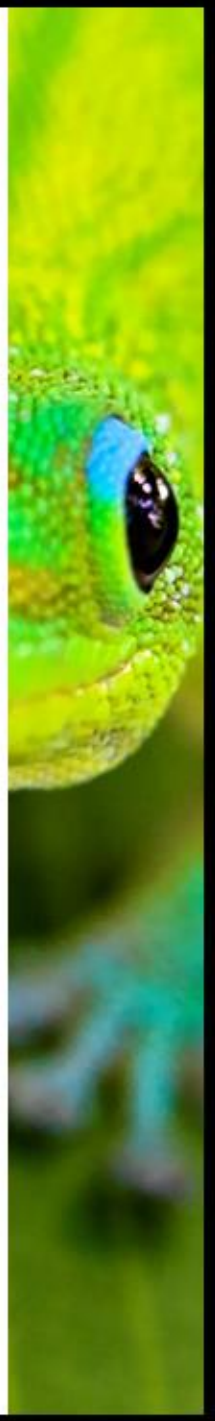
- Liste pré-remplie
- Choix restreints:

Subsp.	Sous-espèce
Var.	Variété
Subvar.	Sous-variété
F.	Forme
Subf.	Sous-forme



Cultivars et hybrides

- Cas complexes et variables: **nécessité d'une base de données sur mesure** !
- Cultivars: **code de nomenclature dédié.**
- Ajouter un flag "cultivar?" et un "hybride?"

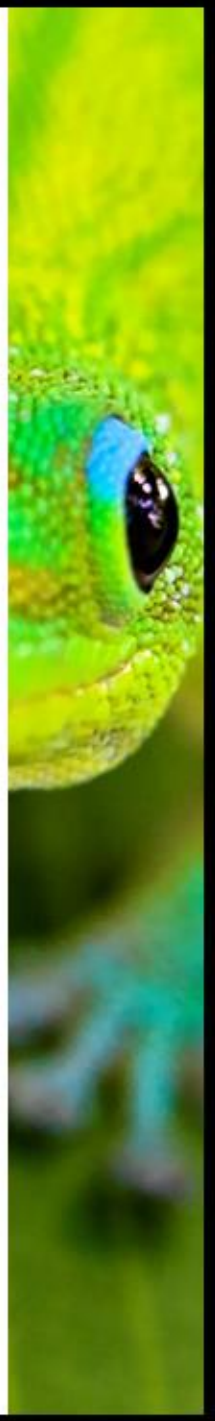


Données taxonomiques

Espèce non publiée – A éviter

- Éviter la confusion avec un nom accepté (pas de nomenclature binomiale pour éviter les erreurs)

- Éviter la confusion entre spécialistes ou institutions (sp1, sp2, …)



Données taxonomiques

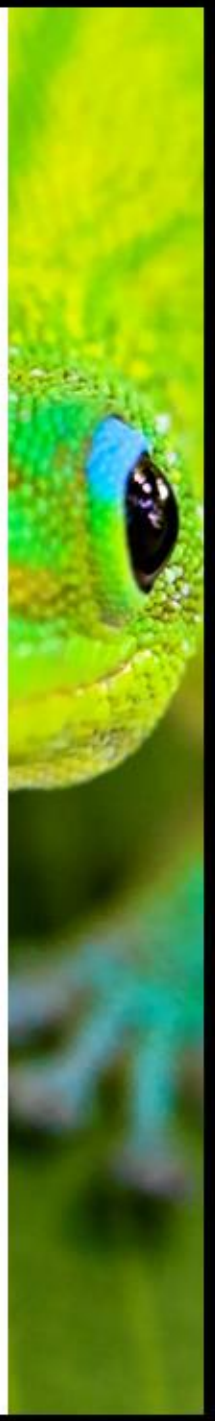
Espèce non publiée – Bonnes pratiques

"<Genus> sp. <colloquial name or description> (<Voucher>)"

Prostanthera sp. Somersbey (B.J. Conn 4024)

Avantages

- Ne ressemble pas à un nom publié
- Pas de confusion entre institutions
- Peut devenir ultérieurement synonyme
- Peu de chances de confusion en dehors du monde scientifique



Données taxonomiques

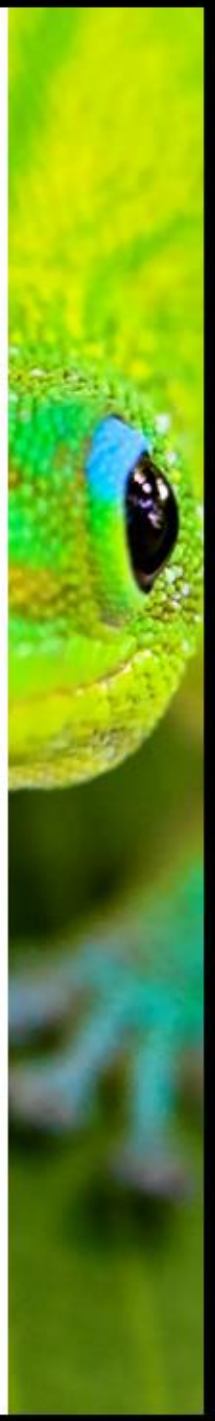
Espèce non publiée – Noms communs

Très complexe à standardiser:

- Un **taxon** = souvent **plusieurs noms**
- Un **nom** = parfois **différents taxons**

Solution: ne pas standardiser (mais **documenter** très largement) !

Nom	Langue	Région	Source	Commentaire



GBIF
www.gbif.org

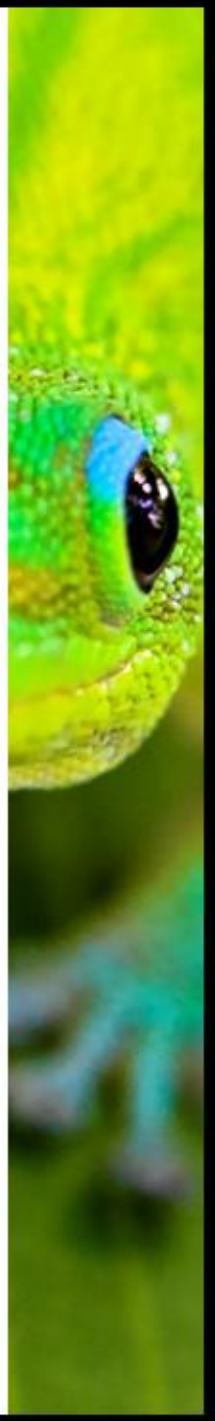
gbif France

Données taxonomiques

Noms des auteurs

- **A renseigner dans des champs séparés**: Genre, espèce, auteur et années

- Pour la nomenclature, tenir compte des **différences entre zoologie** (genre + espèce + auteur + année) et **botanique** (genre + espèce + auteur sans l'année)



Données taxonomiques

Auteur – méthodes de vérification

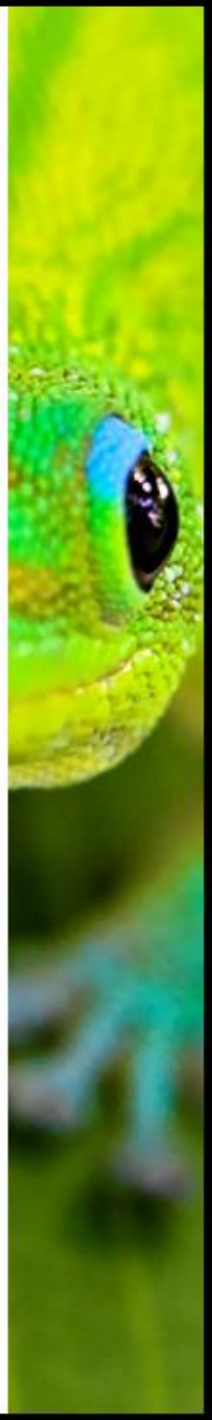
- **Standard pour les abréviations** (plantes) afin d'éviter les doublons
- Fichiers **d'autorité pour l'orthographe (référentiels, bibliographie…)**
- Auteurs **manquants → à compléter**

Données taxonomiques

Nom de collecteur

- La forme doit être standardisée : nom de famille avec initiale en majuscule, initiales en majuscules séparées par des points

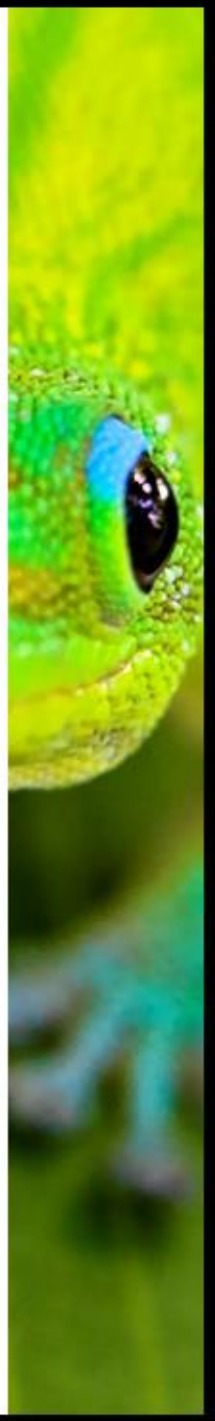
- Ex : Grandidier, A.



Données taxonomiques

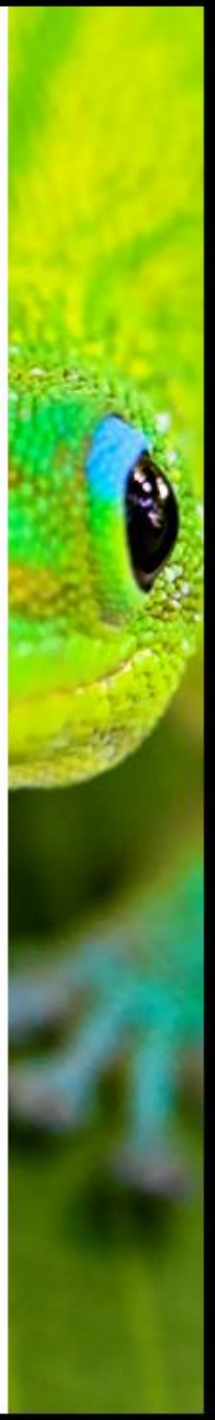
Collecteur: recherche d'erreurs

- Rechercher des variations mineures (voir la démonstration d'Open Refine)
- Comparaisons à d'autres bases: historiques, bibliographiques,…



Bonnes pratiques

Pour les données spatiales

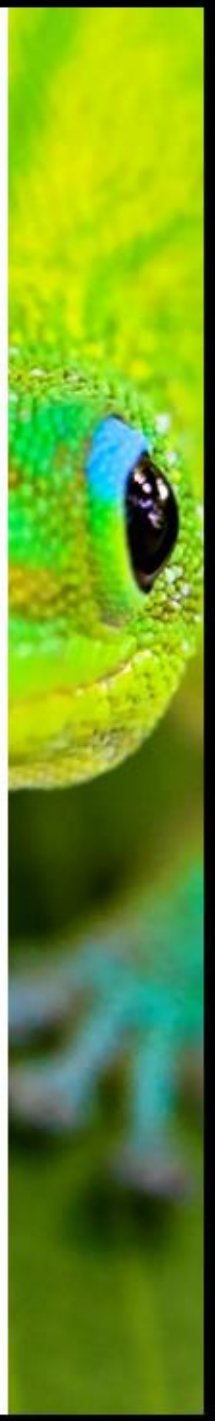


Données spatiales

- Souvent, beaucoup trop de choses dans les champs localité/distribution.

Eurasia: throughout Europe to northernmost extremity of Scandinavia, except Iberian Peninsula, central Italy, and Adriatic basin; Aegean Sea basin in Matriza and from Struma to Aliakmon drainages; Aral Sea basin; Siberia in rivers draining the Arctic Ocean eastward to Kolyma. Widely introduced. Several countries report adverse ecological impact after introduction.

(distribution de *Perca Fluviatilis* selon fishbase)



GBIF
www.gbif.org

gbif France

Données spatiales

Coordonnées décimales (ex: 21.339)

21° 20'20" (DD° MM'SS")

21:20:21

12° 25m

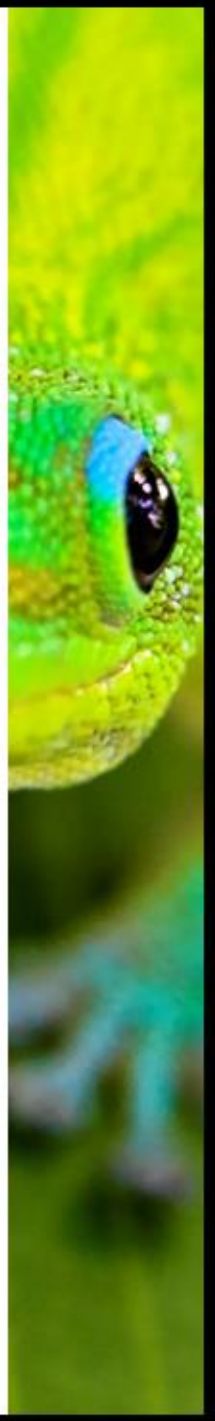
12d25

30' 50" W

North 21 deg 20 min 11,453 sec

N 21 25,568150°

Toujours noter la localité en plus des coordonnées GPS pour confirmer les coordonnées en cas de doute

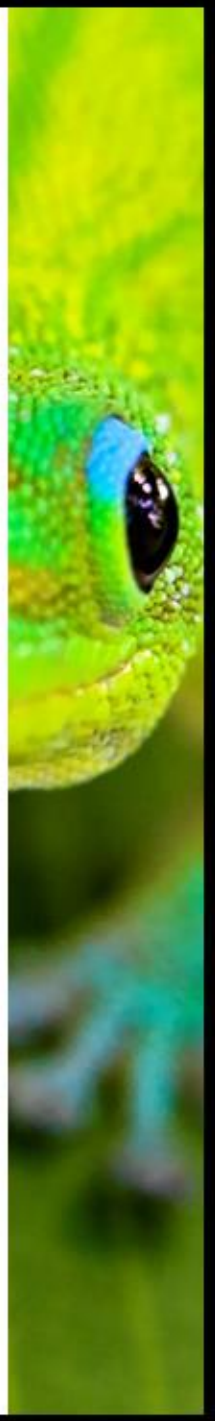


Données spatiales

Datum (type de géoïde + ellipsoïde), système de coordonnées (géographique ou planes) et projection utilisée

SRS (Spatial Reference System/**systèmes de coordonnées géoreférencées**)

Information à documenter!



Données spatiales

Autres informations à fournir :

Précision (rapportée par le GPS): nombre de décimales

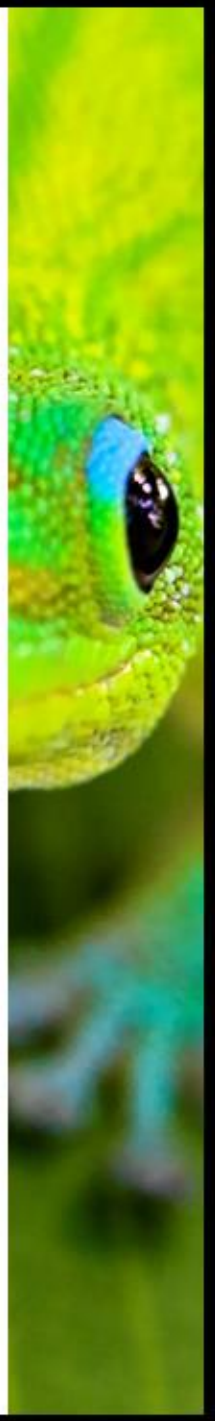
Incertitude spatiale (en mètres si possible): erreurs de géolocalisation (GPS variable de 2 à plus de 20 mètres)
Nom de le lieu plus proche + distance + direction + méthode de géoréférencement

Méthode de géoréférencement

(Differential) GPS: erreur de 10cm a 15m.
'Normal' GPS: erreur de 2 à 20 mètres.

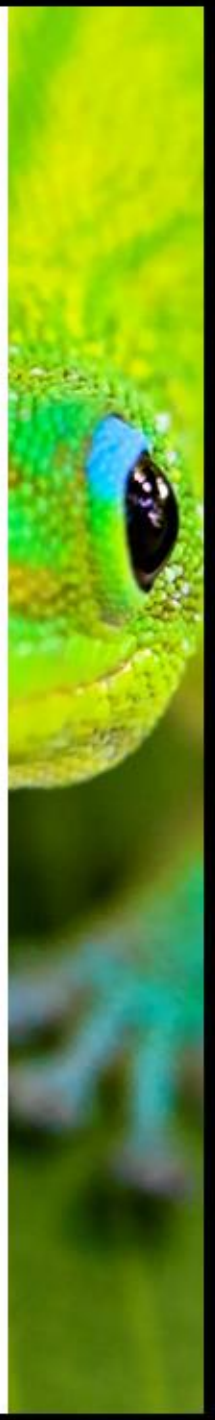
Via carte et triangulation (+échelle)

A posteriori, via un logiciel de géoréférencement (Système d'Information Géographique)



Données spatiales

Détection et correction des erreurs



- Tests **internes**: localité, pays…
- Tests envers des données **externes**: cohérence des noms des lieux visités par le collecteur ? (ex: www.geonames.org pour télécharger base de donnés des noms géographiques; également services web)
- Tests **via un SIG**: test point-dans-polygone ? (terrestre ou marin, pays, régions visités par le collecteur …)
- Recherche de valeurs extrêmes (outliers): **géographiques** ou **environnementales**

Données spatiales

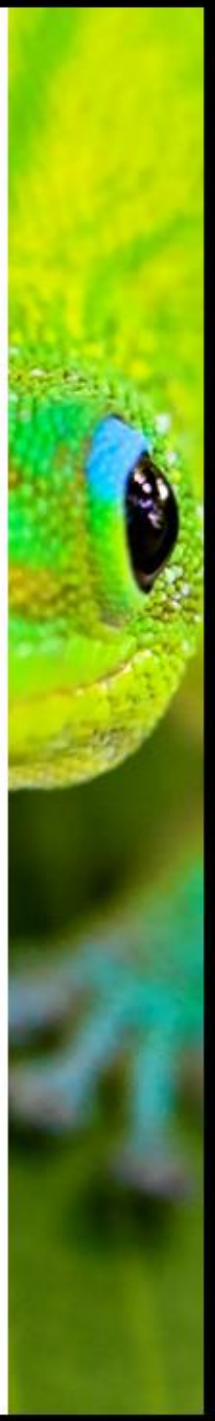
Localité: bonnes pratiques

Noms aussi **spécifiques** que possible:

- Non-ambigus (homonymies, lieux-dits…)
- Courts si possible
- Facile à trouver
- Référence des lieux **stables** et connus
- Distance et direction depuis cette référence

« 2.1km N et 5.1 km E de la la ville de X … »

« A presque 650 mètres de la (petite) rivière Y »

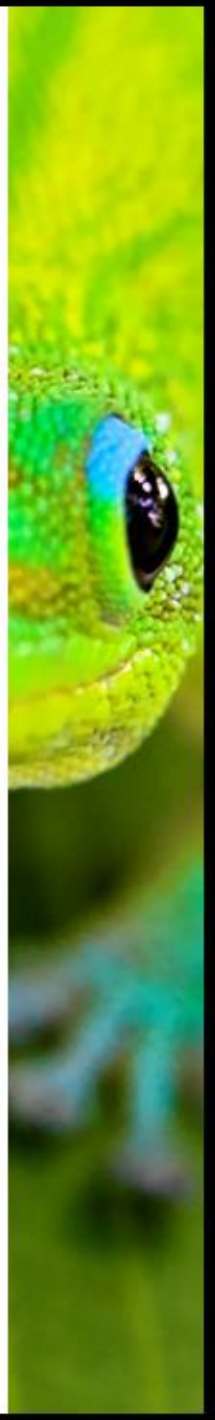


GBIF
www.gbif.org

gbif France

Bonnes pratiques

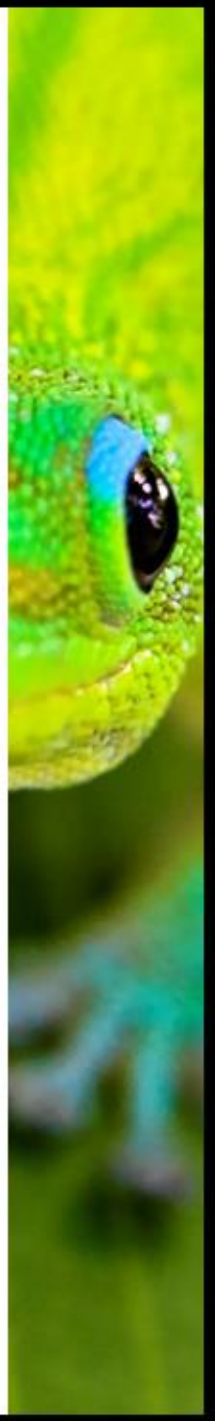
Pour les données sensibles



Données sensibles

Généralisation – pourquoi ?

- **Protéger** les espèces menacées, d'importance économique, réduire l'impact sur les populations sauvages, …
- **Éviter** la collecte non-scrupuleuse, le braconnage, encadrer la bio-prospection,…
- **Protéger les données externes** détenues par l'institution
- **Conserver un avantage compétitif** (publications et recherche)
- **Crainte** d'un usage inapproprié des données
- **Respect**
- **…**



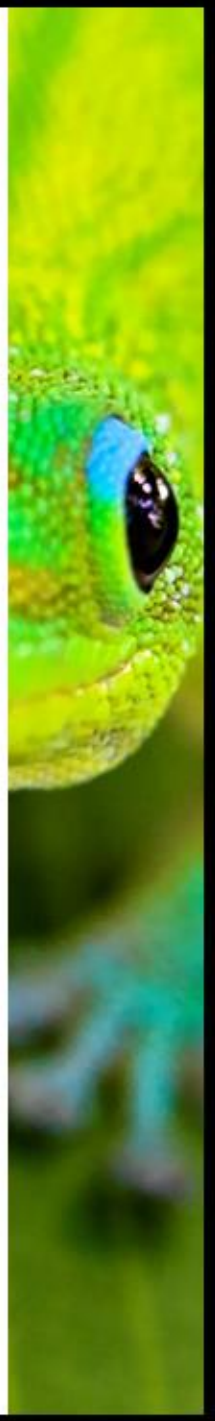
GBIF
www.gbif.org

gbif France

Données sensibles

Généralisation – considérations générales

- **Aspect social** = obstacle principal
- **Composante régionale**
- **Législation** du pays
- La **documentation** est primordiale

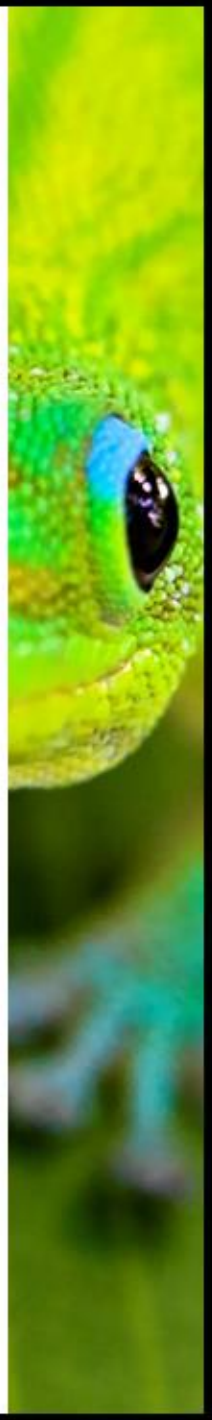


Données sensibles

Généralisation – la doc. est primordiale

Décrire comment et pourquoi les données ont été généralisées permet à l'utilisateur de:

- **Savoir que les données ont été modifiées** et de quelle façon
- **Savoir** qu'il sera **peut-être possible d'obtenir des données plus détaillées**
- Décider **d'ignorer ces données si elle ne conviennent pas à l'usage qu'on veut en faire**, des **les utiliser telles quelles** ou de **chercher des informations supplémentaires**



GBIF
www.gbif.org

gbif France

Données sensibles

Généralisation – comment faire

- **Données spatiales**
  - **Utilisation d'une grille**
  - **3 niveaux recommandés** par Chapman & Wieczorek (2006): 0.1 degrés (11-16 km) - 0.01 degrés (1.1-1.6km) - 0.001 degrés (112-157m)
  - **Cas critiques**: non publiés

- **Données non-spatiales**
  - A remplacer par **une formulation appropriée** (ex : donnée non renseignée pour des raisons légales) afin d'éviter les confusions avec les valeurs « nulles » (non renseignées)
  - Ne **pas restreindre** les **données de collection**

GBIF
www.gbif.org

gbif France

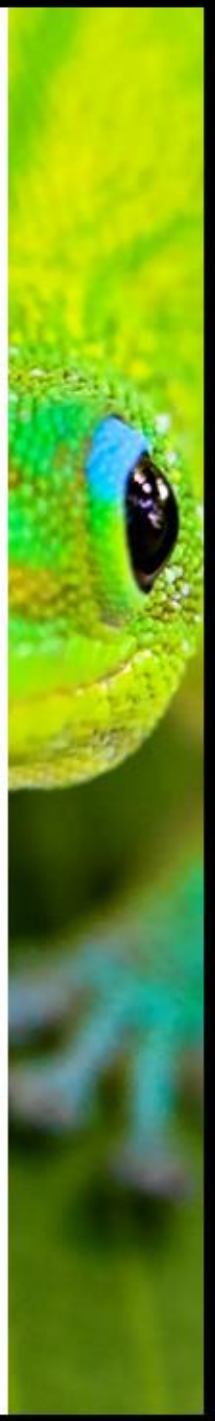
Données sensibles

Généralisation – quoi ?

- **Localité** ou **coordonnées** (cas le plus répandu)

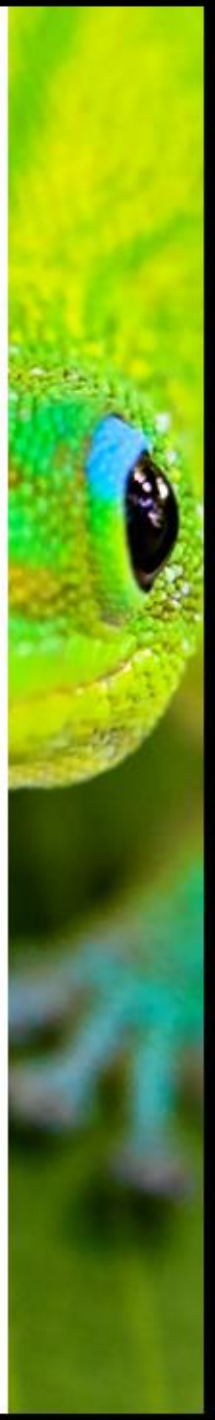
- Autres champs:

informations taxonomiques, identité du collecteur, information sur les habitats, usage traditionnels…



Bonnes pratiques

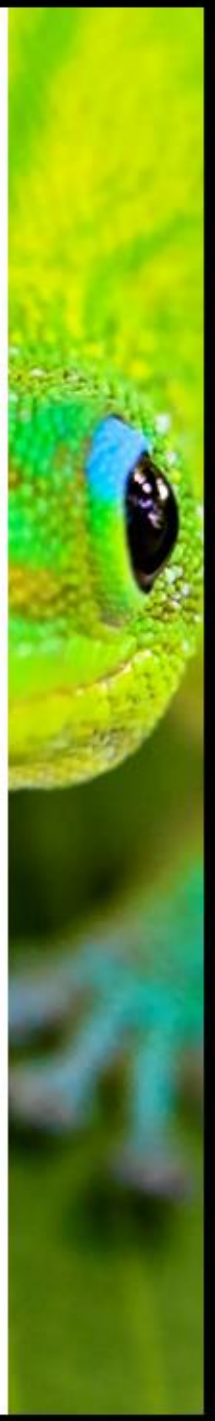
Spécificités GBIF



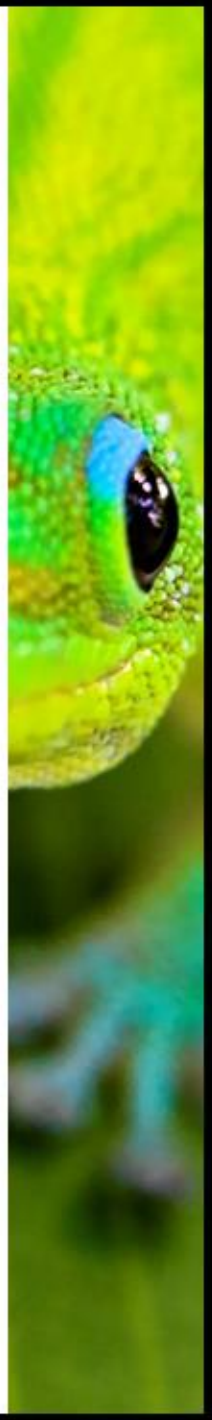
Normalisation GBIF (Darwin Core)

Date – Coordonnées - BasisOfRecord

- **Date**
  - **Format** (ISO 8601:2004(E))
  - Date simple : AAAA-MM-JJ ou AAAA-MM ou AAAA
  - Période : AAAA-MM-JJ/JJ ou AAAA-MM-JJ/MM-JJ ou AAAA/AAAA etc
- **Coordonnées géographiques : lat/long décimales**
- **BasisOfRecord**
  - **Format** Darwin Core Type Vocabulary recommandé
    - PreservedSpecimen
    - FossilSpecimen
    - LivingSpecimen
    - HumanObservation
    - MachineObservation



Pour aller plus loin : outils du GBIF



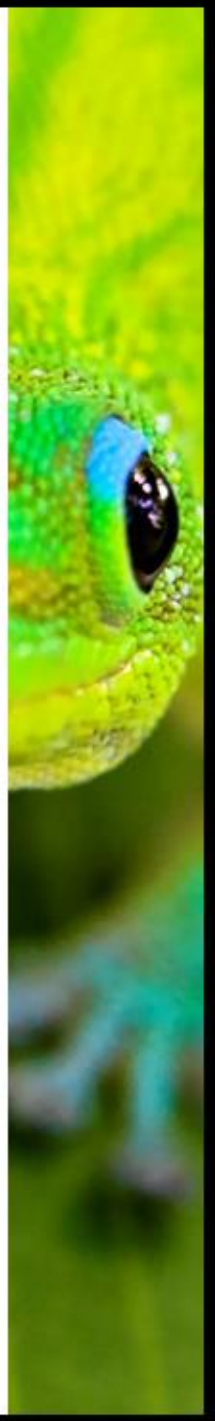
De nombreux outils développés par et pour la communauté GBIF : vérification taxonomique, géographique, …

Liste complète disponible sur le **Biodiversity Data Quality Hub** :

http://www.gbif.es/BDQ.php

Références

Présentation basée sur les publications et les présentations d'Arthur Chapman : « Principles of data quality » et « Principles and methods of data cleaning »



Merci pour votre attention