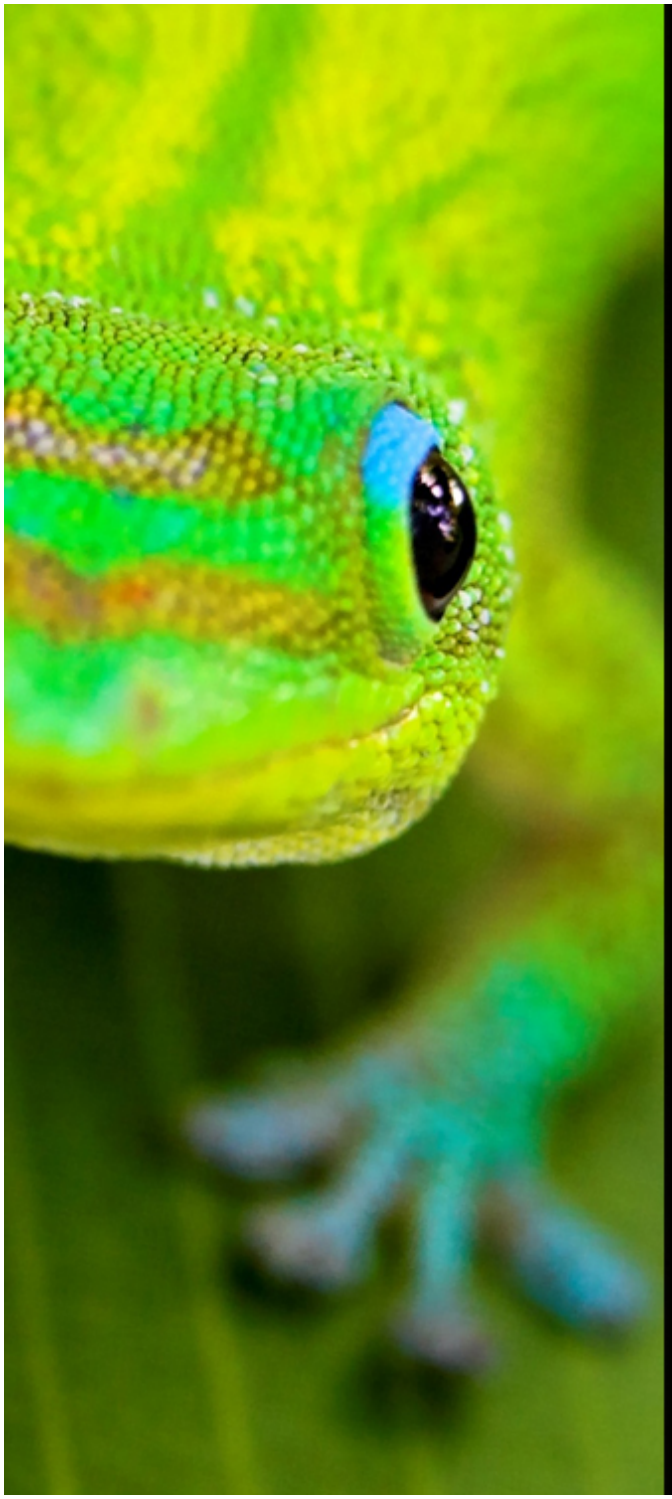


Formation qualité, publication et utilisation des données- Paris, 24-25 mars 2014

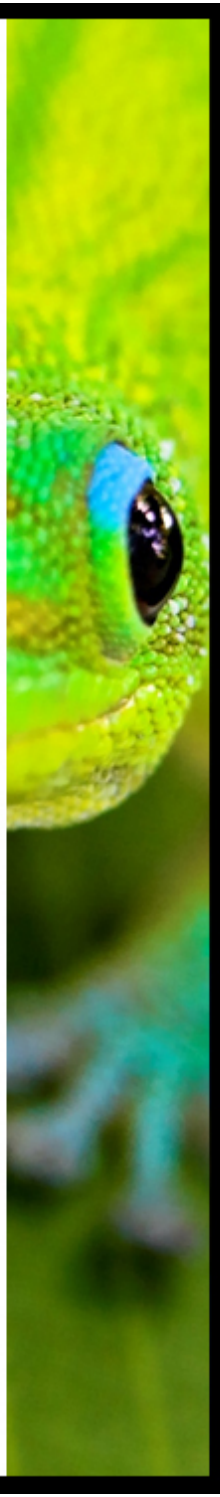
Méthodes et outils pour améliorer la qualité des données de biodiversité

Sophie Pamerlon (pamerlon@gbif.fr)

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

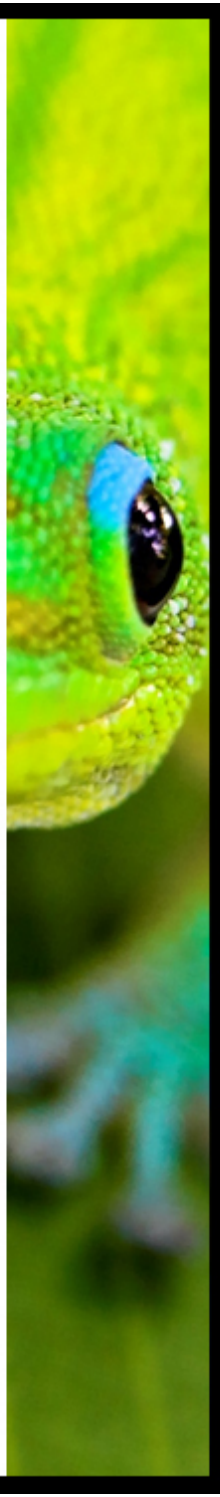
Aperçu

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



Bonnes pratiques

Pour les données taxonomiques

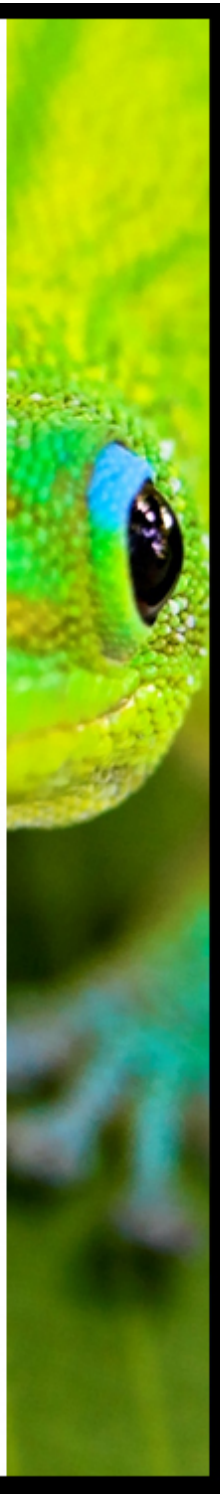


Données taxonomiques

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
  - ...



GBIF
www.gbif.org

gbif France

Données taxonomiques

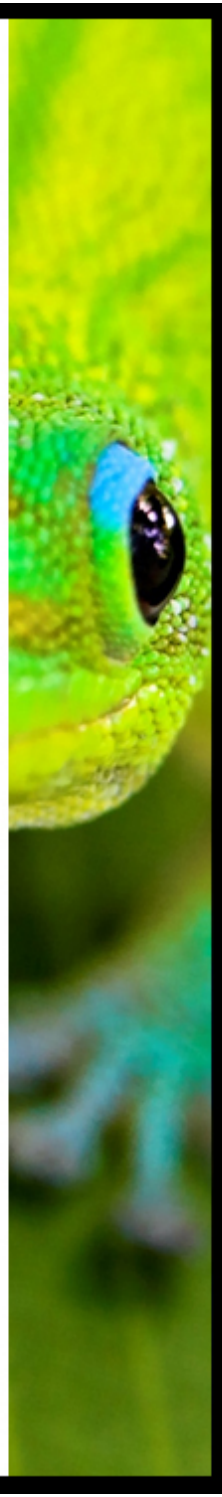
Certitude d'identification

Saisie des données:

- Utilisation de checklists
- Utilisation de fichiers d'autorité

Détection d'erreurs:

- Nécessite généralement un expert
- Les valeurs extrêmes (outliers) peuvent aider (géographiques ou environnementales)



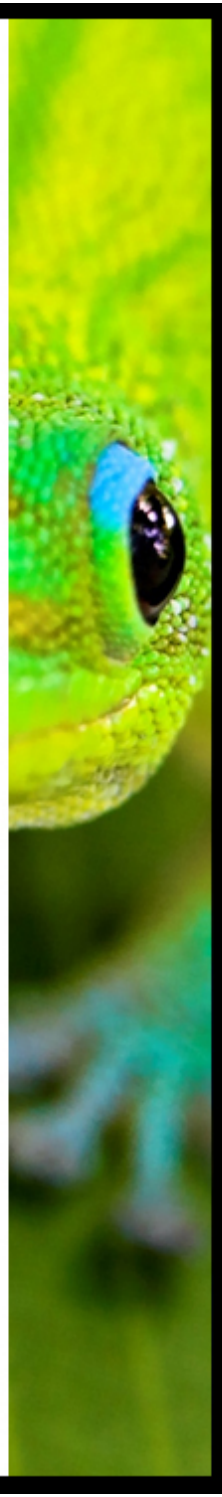
GBIF
www.gbif.org

gbif France

Données taxonomiques

Erreurs orthographiques – nom scientifique

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

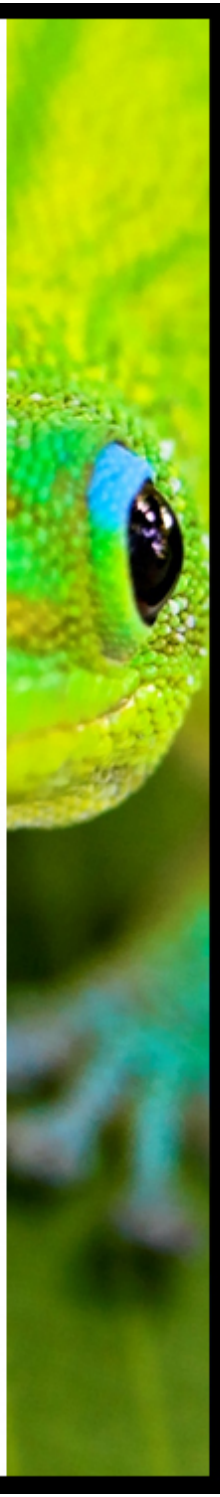


GBIF
www.gbif.org

gbif France

Données taxonomiques

Erreurs orthographiques – rang infra-spécifique



Standardiser !

Genus	Espèce	Rang_infra	Val_infra
Stipiturus	malachurus	Subsp.	parimeda

Pour

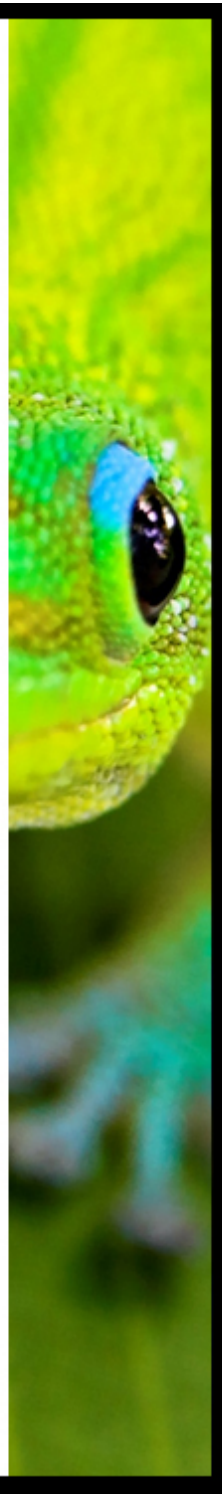
- Éviter les ambiguïtés
- Faciliter les vérifications

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

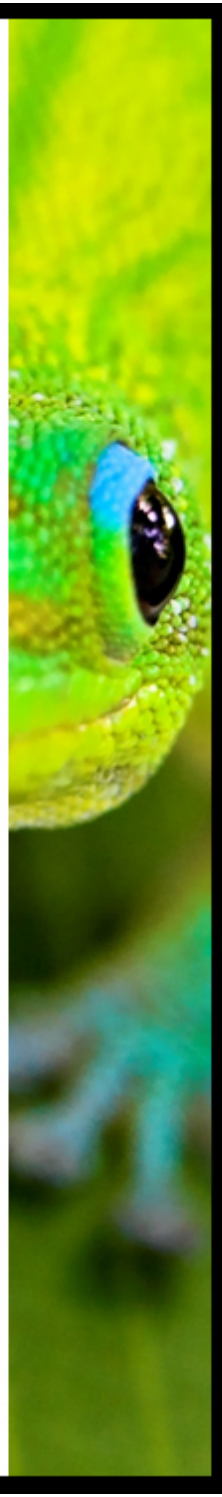


GBIF
www.gbif.org

gbif France

Cultivars et hybrides

- Cas complexes et variables: **DB sur mesure** !
- Cultivars: **code de nomenclature dédié.**
- Ajouter un flag “cultivar?” et un “hybride?”



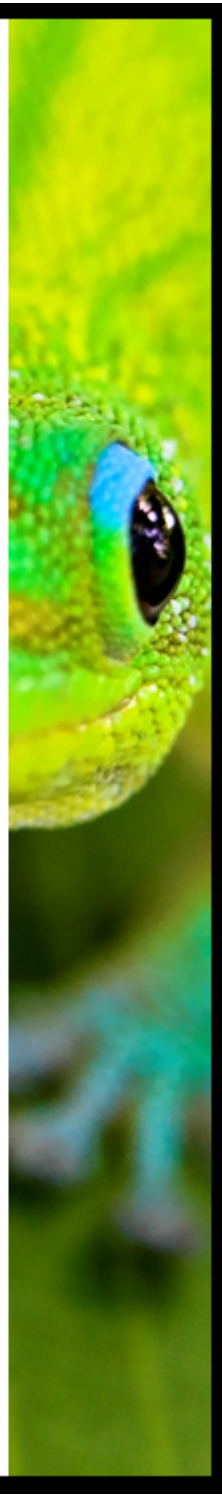
GBIF
www.gbif.org

gbif France

Données taxonomiques

Espèce non publiée – A éviter

- Éviter la confusion avec un nom accepté !
- Éviter la confusion entre spécialistes ou institutions (sp1, sp2, …)



GBIF
www.gbif.org

gbif France

Données taxonomiques

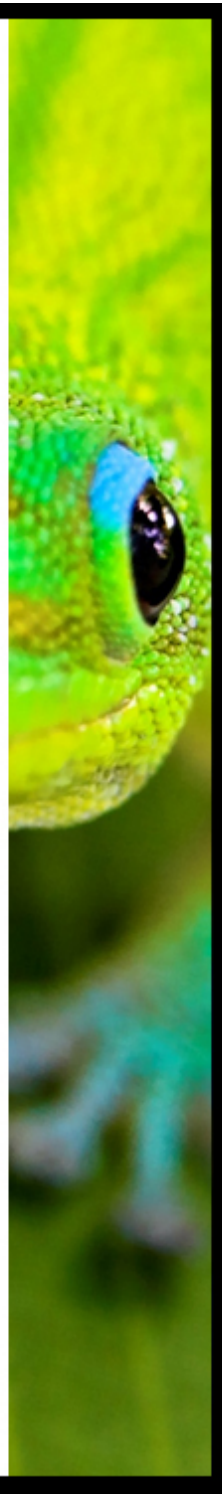
Espèce non publiée – Bonnes pratiques

"<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



GBIF
www.gbif.org

gbif France

Données taxonomiques

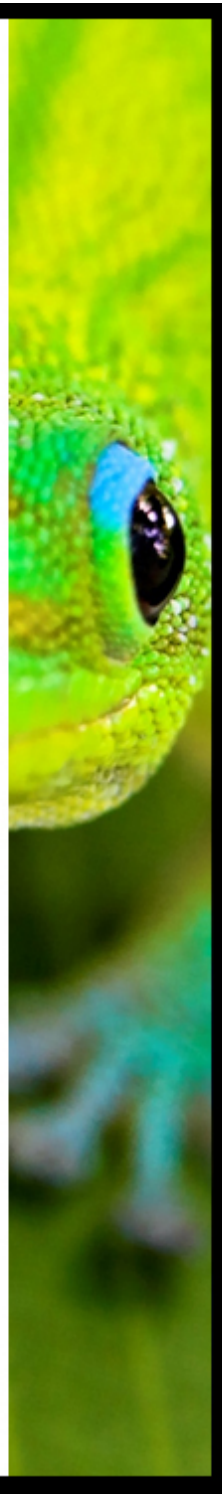
Espèce non publiée – Noms communs

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



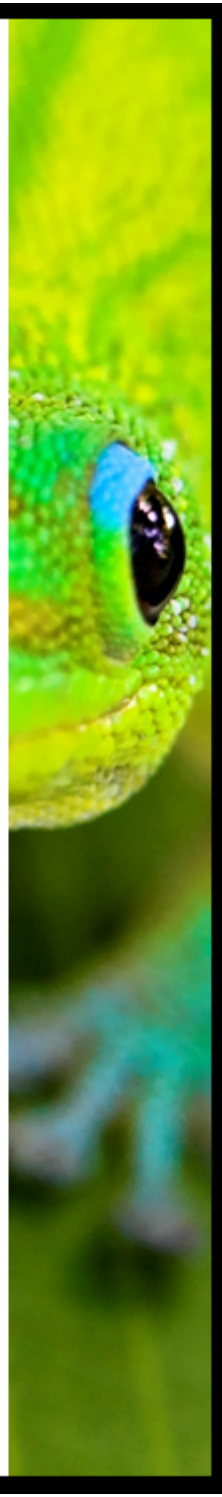
GBIF
www.gbif.org

gbif France

Données taxonomiques

Noms des auteurs

- **Rarement vraiment nécessaire**
- Si inclus: **champs séparés**: Genre, espèce, auteur et années
- Pour l'affichage, tenir compte des **différences entre animaux et végétaux**



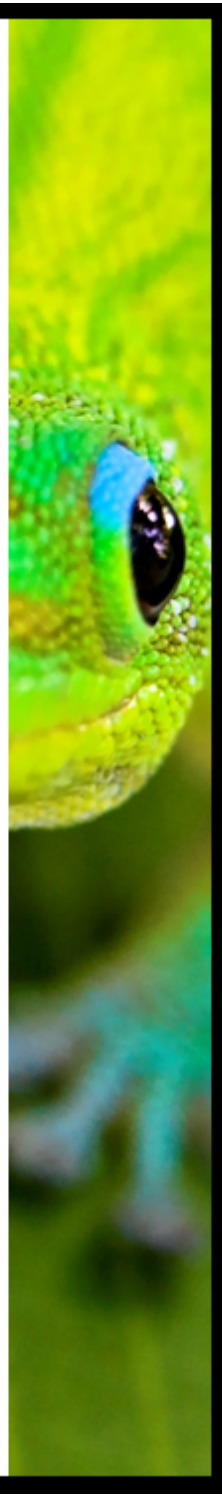
GBIF
www.gbif.org

gbif France

Données taxonomiques

Auteur – méthodes de vérification

- **Standard pour les abréviations** (plantes)
- Fichiers **d'autorité**
- **Soundex**
- Auteurs **manquants**

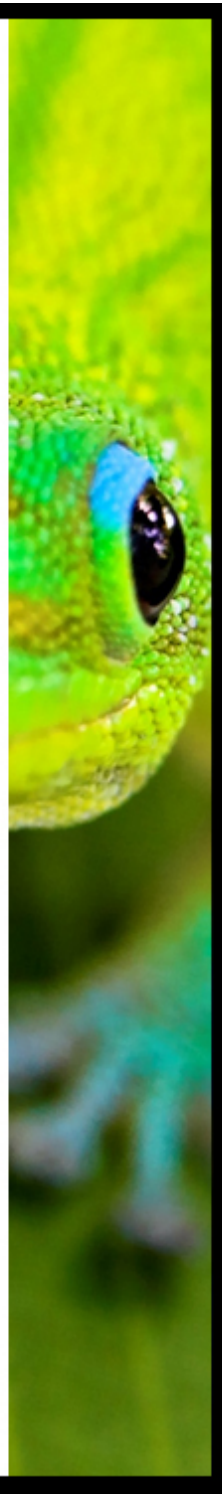


Données taxonomiques

Nom de collecteur

- Parfois, liste exhaustive
- La forme doit être standardisée

"Primary collector's family name (surname) followed by comma and space (,) then initials (all in uppercase and each separated by fullstops). All initials and first letter of the collector's family name in uppercase. For example, Chambers, P.F."



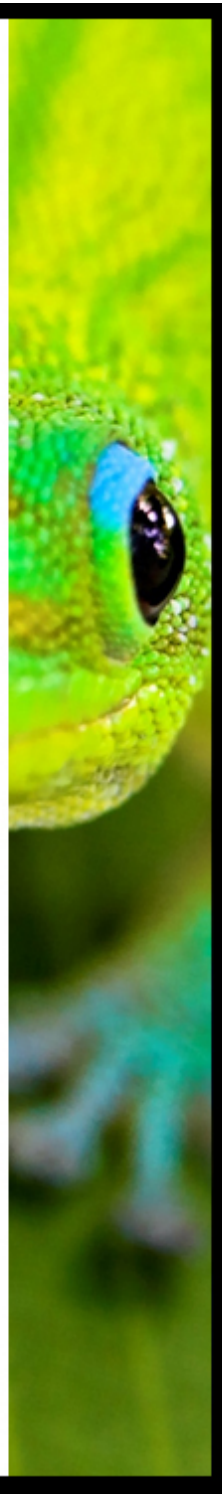
GBIF
www.gbif.org

gbif France

Données taxonomiques

Collecteur: recherche d'erreurs

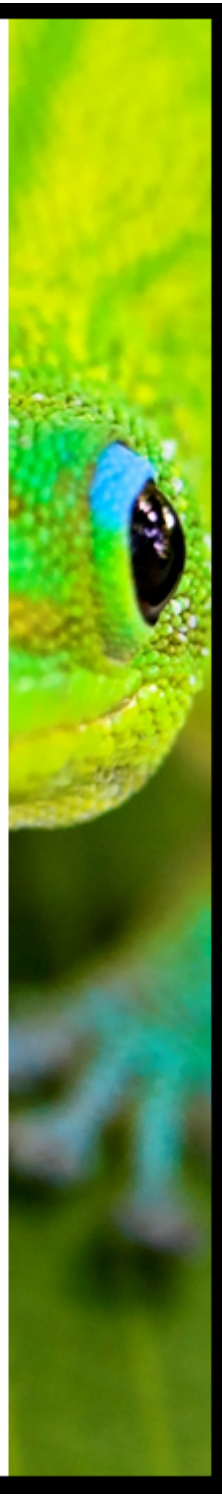
- Rechercher des variations mineures
- Comparaisons à d'autres bases: historiques, …



Améliorations possibles dans les deux sens !

Bonnes pratiques

Pour les données spatiales

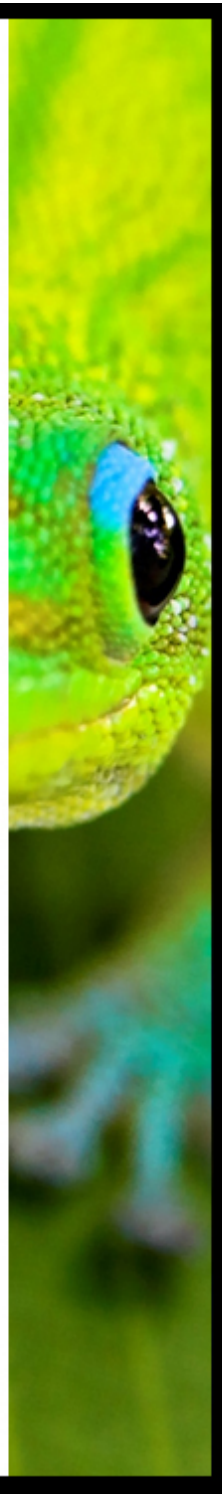


Données spatiales

- 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

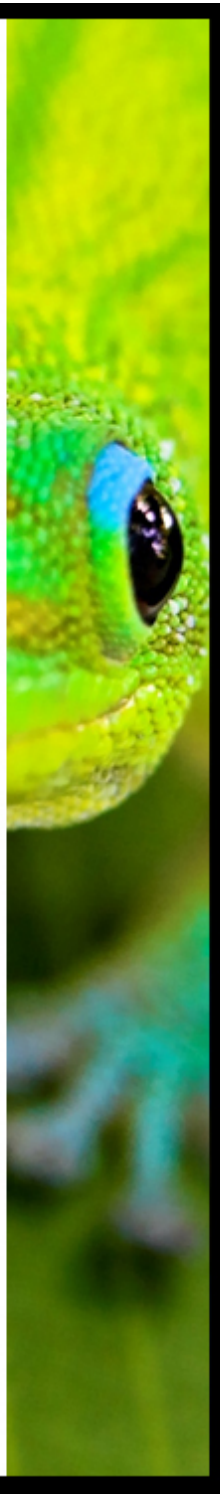
30' 50" W

North 21 deg 20 min 11,453 sec

N 21 25,568150°

Outil de conversion on-line:

http://webapps.cartoninjas.net/coordinateconverter/



GBIF
www.gbif.org

gbif France

Données spatiales

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

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

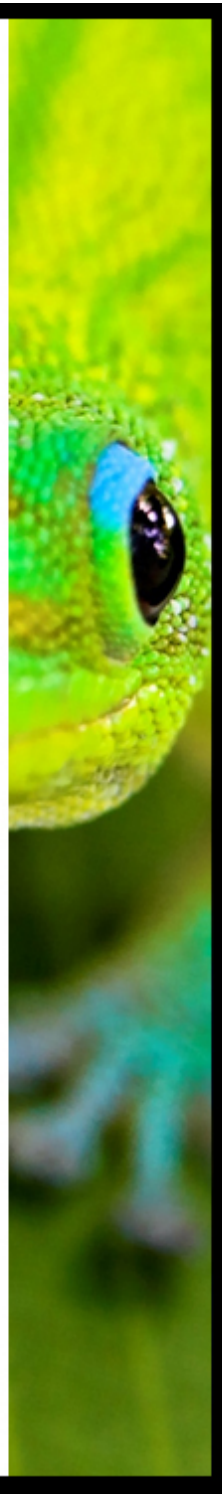
Information à documenter!

```
PROJCS["NAD27(76) / UTM zone 17N",
    GEOGCS["NAD27(76)",
        DATUM["North_American_Datum_1927_1976",
            SPHEROID["Clarke 1866",6378206.4,294.9786982138982,
                AUTHORITY["EPSG","7008"]],
            AUTHORITY["EPSG","6608"]],
        PRIMEM["Greenwich",0,
            AUTHORITY["EPSG","8901"]],
        UNIT["degree",0.01745329251994328,
            AUTHORITY["EPSG","9122"]],
        AUTHORITY["EPSG","4608"]],
    UNIT["metre",1,
        AUTHORITY["EPSG","9001"]],
    PROJECTION["Transverse_Mercator"],
    PARAMETER["latitude_of_origin",0],
    PARAMETER["central_meridian",-81],
    PARAMETER["scale_factor",0.9996],
    PARAMETER["false_easting",500000],
    PARAMETER["false_northing",0],
    AUTHORITY["EPSG","2029"],
    AXIS["Easting",EAST],
    AXIS["Northing",NORTH]]
```

UTM, zone 17, Datum NAD27

code 2029 (code EPSG, European Petroleum Survey Group)

Le DATUM ou code EPSG suffit



GBIF
www.gbif.org

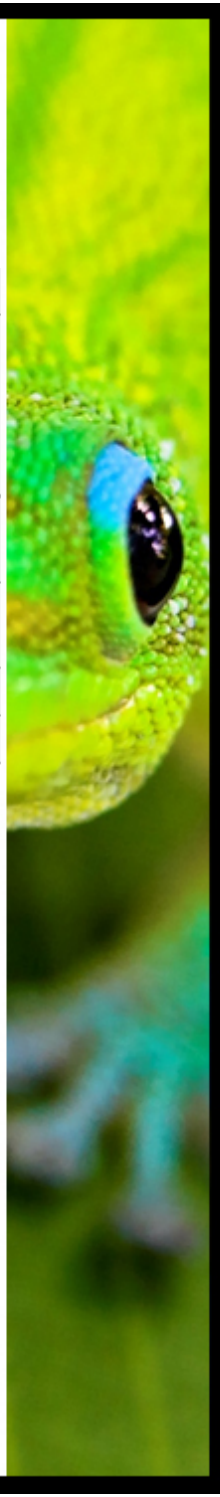
gbif France

Données spatiales

Code	nom	EPSG	Remarques
RGF93	Réseau Géodésique Français 1993	6171 (système géocentrique), 4965 (3D), 4171 (2D)	Système français légal (décret 2000-1276 du 26 décembre 2000). Identique à l'ETRS89 au 1/1/1993. Compatible avec le WGS84 pour des précisions égales ou supérieures à 10 m (c'est-à-dire 15 m etc.).
NTF	Nouvelle Triangulation Française	2D : 4807 (Paris, grade) ou 4275 (Greenwich, degré). 3D : 7400 (Paris, grade)	Système français périmé mais encore largement utilisé.
ETRS89	European Terrestrial Reference System 1989	4937 (3D), 4258(2D)	Système européen actuel
ED50	European Datum 1950	4230	Système européen périmé
WGS84	World Geodetic System 1984	4979 (3D), 4326 (2D)	Système mondial très utilisé notamment avec le GPS.

GBIF Darwin Core geodeticDatum :

http://rs.tdwg.org/dwc/terms/#geodeticDatum



GBIF
www.gbif.org

gbif France

Données spatiales

Autres informations à fournir :

Précision (rapportée par le GPS): nombre de décimales

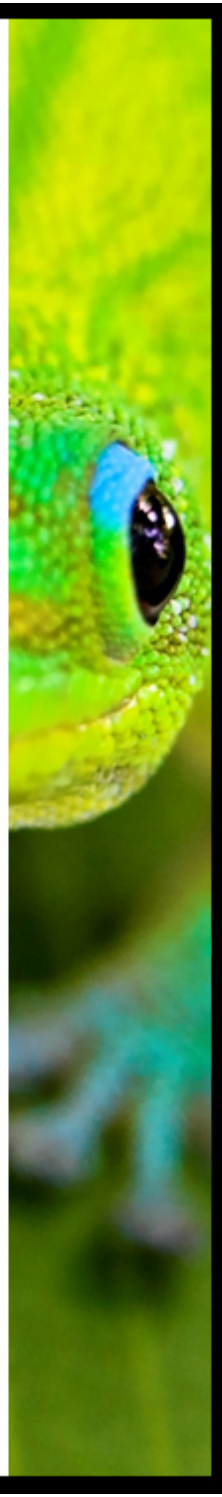
Incertitude spatiale (en metres si possible): erreurs de géolocalisation (GPS variable de 2 à plus de 20 mètres)
Nom de le lieu plus proche + distance + direction + méthode de géoréférencement

Méthode de géoréférencement

(Differential) GPS: erreur de 10cm a 15m.
'Normal' GPS: erreur de 2 à 20 mètres.
GPS dégradé par « Selective availability » (avant année 2000)

Via carte et triangulation (+échelle)

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



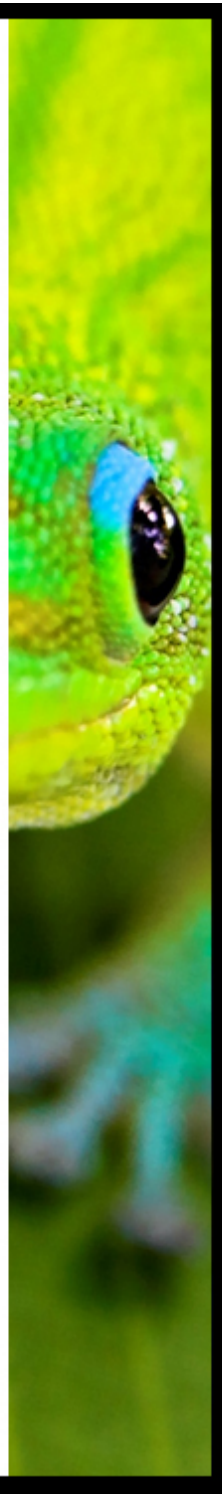
GBIF
www.gbif.org

gbif France

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: Geonames.org pour télécharger base de donnés du 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**



GBIF
www.gbif.org

gbif France

Données spatiales

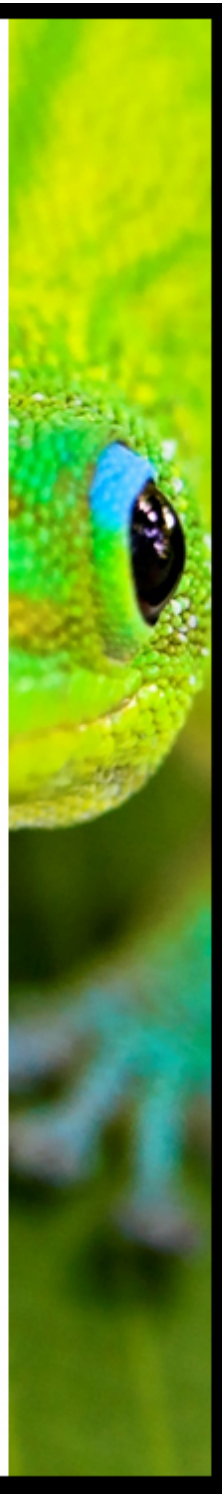
Localité: bonnes pratiques

Aussi **spécifiques** que possible:

- Non-ambiguës
- Courtes
- 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 »

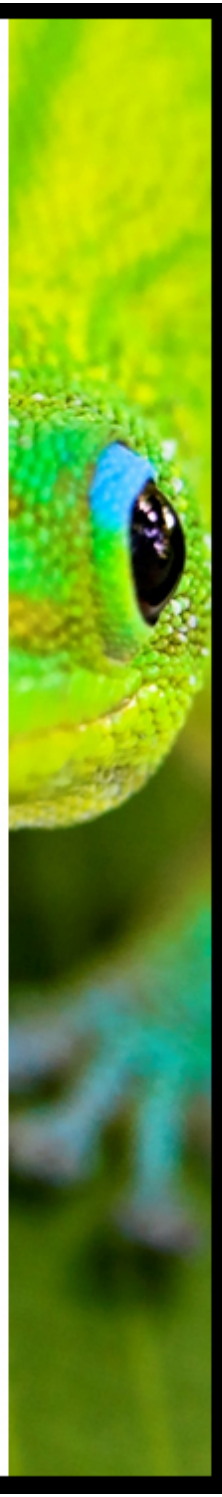


GBIF
www.gbif.org

gbif France

Bonnes pratiques

Pour les données sensibles

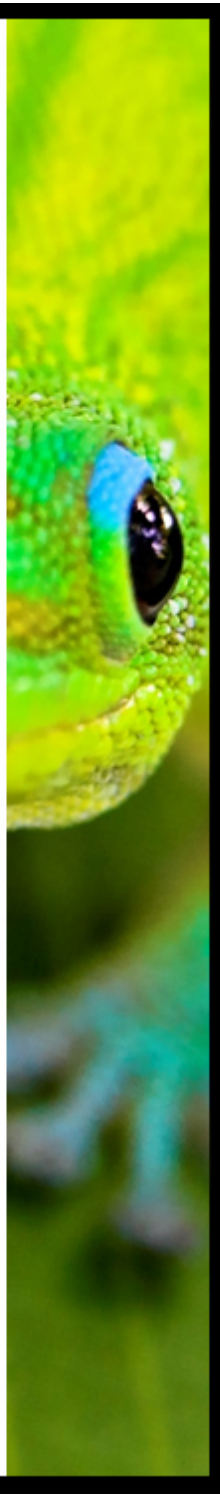


Données sensibles

Généralisation – pourquoi ?

- **Protéger** 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**
- **…**

Résultats du sondage en ligne du GBIF (2006)



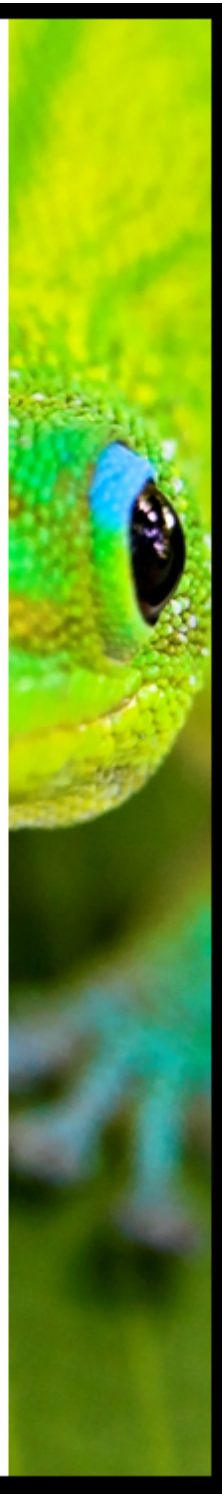
GBIF
www.gbif.org

gbif France

Données sensibles

Généralisation – considérations générales

- **Aspect social** = obstacle principal
- **Composante régionale**
- Certains ne publieront **jamais**
- La **documentation** est primordiale

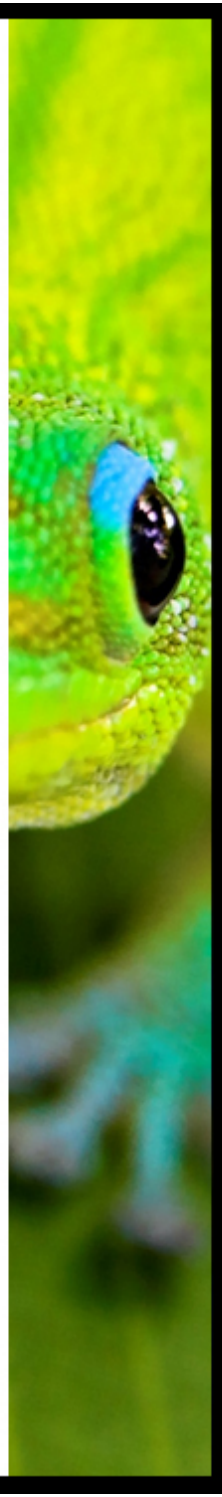


Données sensibles

Généralisation – la doc. est primordiale

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

- **Savoir que les données on été modifiées** et en quoi
- **Savoir** qu'il sera **peut-être possible d'obtenir des données plus détaillées**
- Décider **d'ignorer ces données**, des **les utiliser telles quelles** ou de **chercher des informations supplémentaires**



GBIF
www.gbif.org

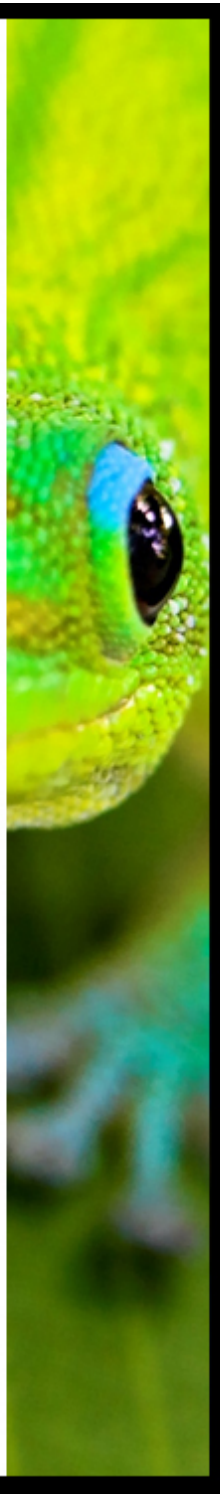
gbif France

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**
  - Ne **pas restreindre** les **données de collection**



GBIF
www.gbif.org

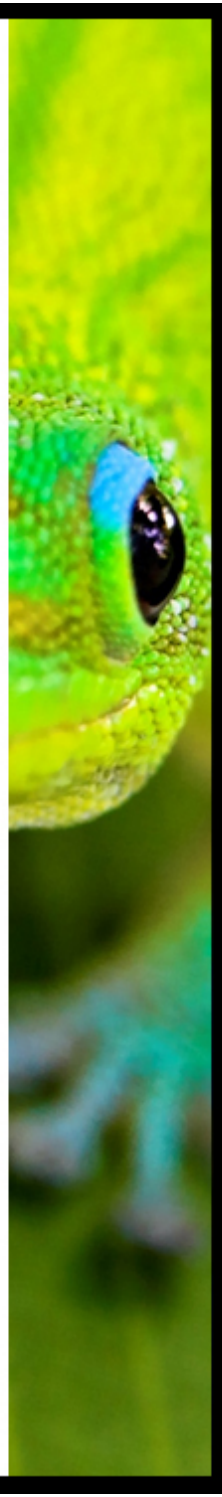
gbif France

Données sensibles

Généralisation – quoi ?

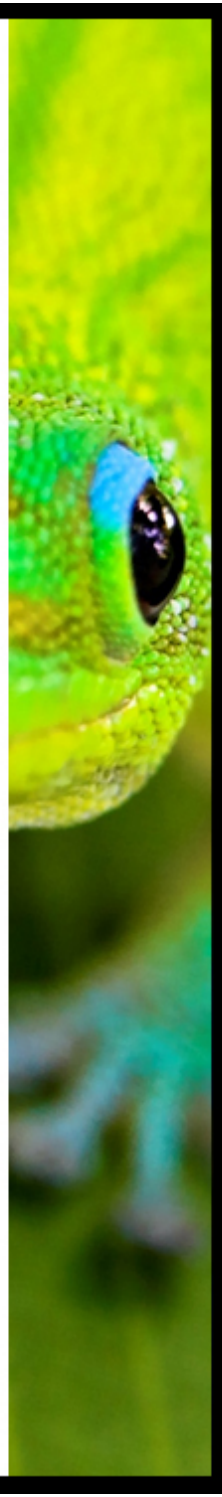
- **Localité** ou **coordonnées**

- Autres champs: informations taxonomiques, identité du collecteur, information sur les habitats, usage traditionnels



Bonnes pratiques

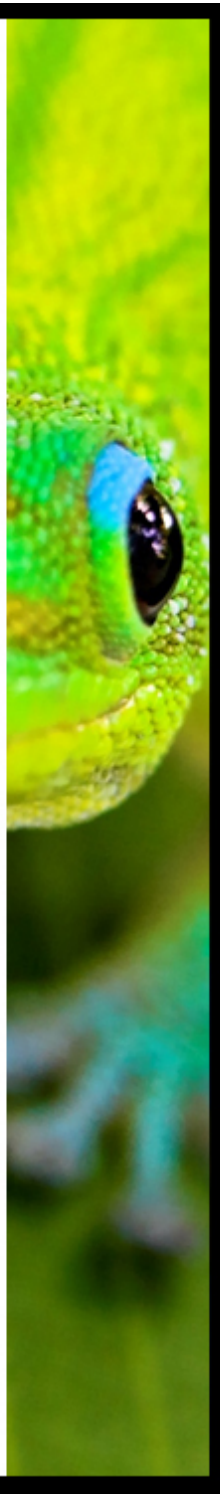
Spécificités GBIF



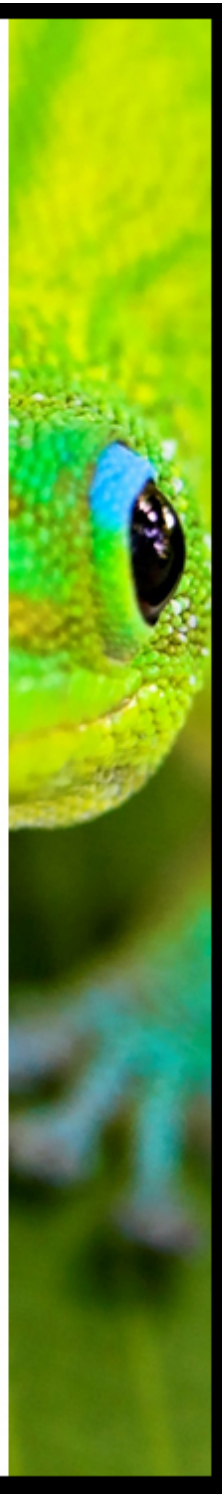
Normalisation GBIF

Date - 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
- **BasisOfRecord**
  - **Format** Darwin Core Type Vocabulary recommandé
    - PreservedSpecimen
    - FossilSpecimen
    - LivingSpecimen
    - HumanObservation
    - MachineObservation



Pour aller plus loin : outils du GBIF



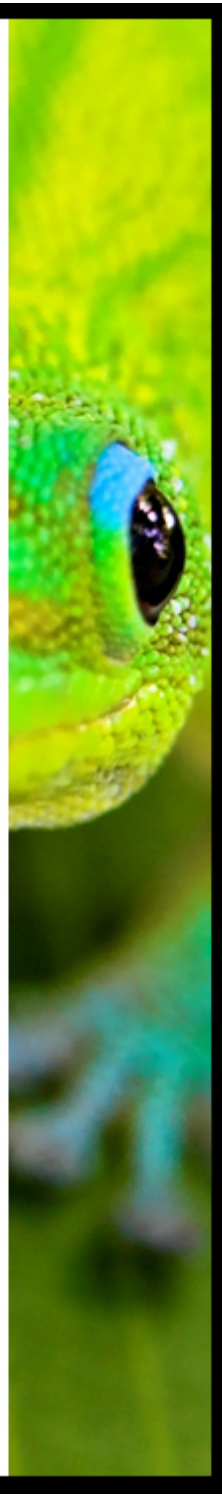
De nombreux outils développés par et pour la communauté GBIF : vérification taxonomique, géographique, …

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

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

Questions ?

Merci



GBIF
www.gbif.org

gbif France

Références

Basé sur les publications et les présentations d'Arthur Chapman :

« Principles of data quality » et

« Principles and methods of data cleaning ».