



**Améliorer la qualité  
des données en  
Nouvelle-Aquitaine :**  
production, contrôles, validation



## Améliorer la qualité des données en Nouvelle-Aquitaine :

- 1 - Lors de leur production
- 2 - Contrôles de conformité  
et de cohérence
- 3 - Validation scientifique
- 4 - Précision et complétude



# Améliorer la qualité des données en Nouvelle-Aquitaine : production, contrôle, validation

## Organisation du SINP en Nouvelle-Aquitaine :

### Co-pilotes du SINP en région



### Plateforme régionale Nouvelle-Aquitaine

(= 3 pôles thématiques)

« faune »



« flore, fonge, habitats »



« géologie »



**Equipe d'animation régionale =** Co-pilotes + Plateforme régionale + Réseau associatif





## Améliorer la qualité des données en Nouvelle-Aquitaine :

### 1 - Lors de leur production



# 1 - Améliorer la qualité des données : lors de leur production

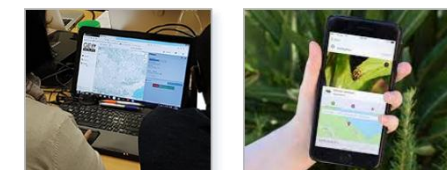
## De multiples sources de données :

- Un **réseau d'acteurs varié** provenant de structures diverses

- Des **méthodes de collecte, de détermination, et de numérisation hétérogènes**



- De **nombreux outils de saisie** de données



- Des **habitudes personnelles**



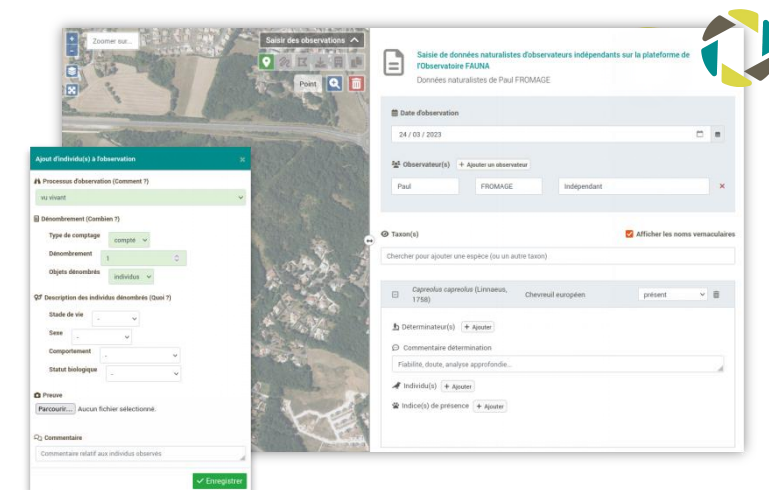
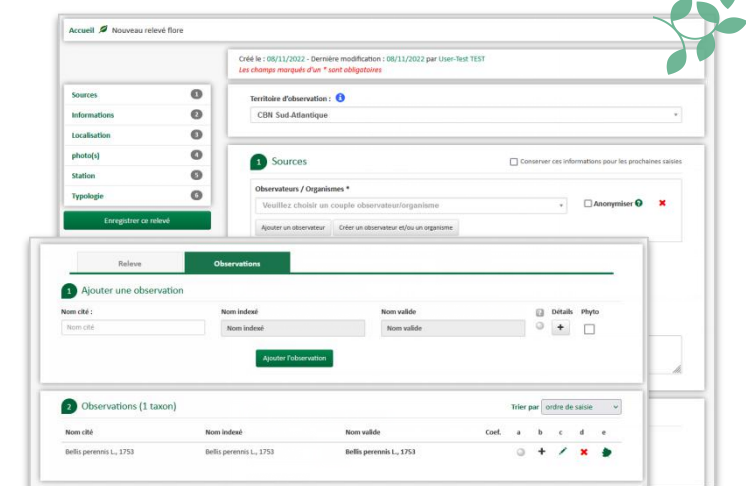




# 1 - Améliorer la qualité des données : lors de leur production

... par la mise à disposition d'outils :

- Interfaces de saisie et de gestion des données et métadonnées sur les sites de FAUNA et de l'OBV-NA
  - **standardisation automatique au format SINP** : données homogènes et interoperables
  - **respect des nomenclatures et référentiels** : TAXREF, référentiels habitat/végétation, stade de vie...
  - **contrôles en direct** : localisation, cohérence (espèce/nomenclature, dates min/max), champs obligatoires...
  - **alertes vers l'utilisateur à propos d'informations importantes à renseigner** (ex. critères pour éléments sensibles)
  - **formulaire thématiques** : adaptation des champs et des nomenclatures proposés par type de relevé (flore/phytosociologique), ou par protocole/projet (suivi des collisions routières, enquête scientifique espèce-centrée...)







## 1 - Améliorer la qualité des données : lors de leur production

### ... par l'accompagnement et la sensibilisation des producteurs et fournisseurs de données :

- en particulier lors de la **création de métadonnées** (*obligatoire pour l'intégration dans le SINP*)
  - pour la définition des cadres d'acquisition et jeux de données
  - contrôle et validation des métadonnées produites : complétude des informations renseignées (*description du protocole, indication de l'ensemble des acteurs impliqués, source des données...*)
- **prévention à la création de doublon** : vérification auprès des fournisseurs de données pour éviter les doubles dépôts (*au niveau national/régional ou par les différents acteurs du projet*)



### ... par la traçabilité des données et métadonnées :

- **identifiant SINP unique permanent** géré en base de données et communiqué à tous
- couple **Base de données / ID source** permettant de remonter à la donnée chez le producteur (*s'il ne dispose pas d'UUID considéré comme SINP*)





## 1 - Améliorer la qualité des données : lors de leur production

Attention, la qualité lors de la production, c'est aussi avant la phase de saisie par :

- les protocoles de collecte mis en place
- les méthodes d'inventaires choisies (*par ex. relevés de surface réduite*)
- la détermination rigoureuse des espèces (*qualité des ouvrages de détermination utilisés, etc..*)
- la collecte et le renseignement d'informations complémentaires aux observations d'espèces (*statut de spontanéité, comportement de reproduction, etc..*)





**Améliorer la qualité  
des données en  
Nouvelle-Aquitaine :**

**2 - Contrôles de conformité  
et de cohérence**





## 2 - Améliorer la qualité des données : contrôles de conformité et de cohérence

Cas des lots de données fournies à la plateforme régionale pour intégration (*non saisies via les interfaces en ligne*)

*soit + 95 % des données du SINP faune*

*et 22 % des données de l'OBV-NA au total (et 37 % des données issues du réseau naturaliste)*



« La **conformité** désigne le **respect des règles fixées** dans le cadre de la mise en œuvre des **formats standards** de données et de métadonnées autant sur les aspects physiques que conceptuels (renseignement des champs obligatoires, format, utilisation des référentiels et des listes de valeurs/nomenclatures). »

« La **cohérence** désigne le **respect de la logique combinatoire des informations** transmises au sein des données, au sein des métadonnées et entre les données et les métadonnées. »

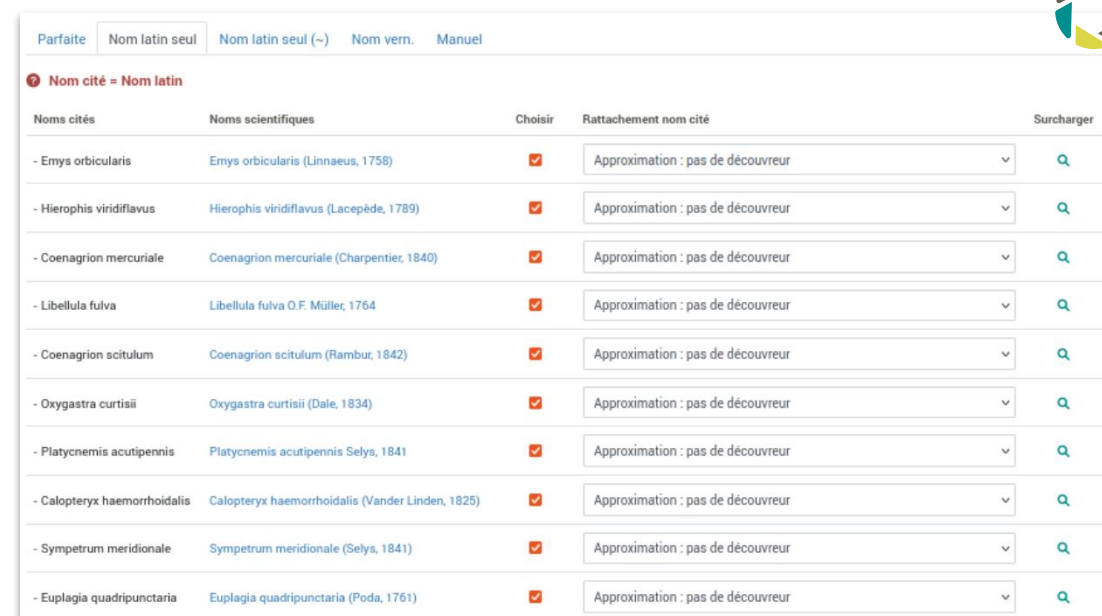
*Exemple : la date de début de l'observation est inférieure ou égale à la date de fin de l'observation.*



## 2 - Améliorer la qualité des données : contrôles de conformité et de cohérence

### Contrôle, formatage et insertion des lots de données reçus :

- analyse des fichiers reçus (étape semi-manuelle)
- rattachement à TAXREF :
  - effectué par la plateforme régionale
  - gestion des versions (avec mise à jour des rattachements lors d'une montée de version)
  - gestion des complexes (regroupements d'espèces)
  - FAUNA : matching TAXREF
- rattachement aux observateurs et organismes
- standardisation selon un format pivot :
  - insertion automatique
  - contrôles de cohérence / conformité (localisation, listes valeurs...)
  - blocage de l'intégration si données non conformes (rattrapage automatique dans certains cas pour la faune via un dictionnaire de synonyme)



Noms cités	Noms scientifiques	Choisir	Rattachement nom cité	Surcharger
- Emys orbicularis	Emys orbicularis (Linnaeus, 1758)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Hierophis viridiflavus	Hierophis viridiflavus (Lacépède, 1789)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Coenagrion mercuriale	Coenagrion mercuriale (Charpentier, 1840)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Libellula fulva	Libellula fulva O.F. Müller, 1764	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Coenagrion scitulum	Coenagrion scitulum (Rambur, 1842)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Oxygastra curtisii	Oxygastra curtisii (Dale, 1834)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Platynemesis acutipennis	Platynemesis acutipennis Selys, 1841	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Calopteryx haemorrhoidalis	Calopteryx haemorrhoidalis (Vander Linden, 1825)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Sympetrum meridionale	Sympetrum meridionale (Selys, 1841)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>
- Euplagia quadripunctaria	Euplagia quadripunctaria (Poda, 1761)	<input checked="" type="checkbox"/>	Approximation : pas de découvreur	<a href="#">Q</a>





**Améliorer la qualité  
des données en  
Nouvelle-Aquitaine :**

**3 - Validation scientifique**



### 3 - Améliorer la qualité des données : validation scientifique

« La **validation scientifique** consiste en des processus d'expertise visant à **renseigner sur la fiabilité** (désigne le degré de confiance que l'on peut accorder à la donnée). Ces processus font intervenir des bases de connaissance et/ou de l'expertise directe. »

Trois niveaux de validation dans le SINP :

- validation producteur
- **validation régionale**
- validation nationale

échange de ces valeurs  
entre plateformes

Différents niveaux de fiabilité d'une observation:

- Certain - Très probable
- Probable
- Douteux
- Invalide - Très douteux
- Non applicable

On distingue :

- la validation **automatique**
- la validation **manuelle**

Combinaison de ces deux méthodes pour la validation régionale en Nouvelle-Aquitaine





### 3 - Améliorer la qualité des données : validation scientifique

#### Combinaison des méthodes de validation automatique et manuelle :

- validation automatique calculée dès l'insertion pour certains groupes :
  - flore vasculaire (selon une liste de taxons pré-établie)
  - vertébrés
  - quelques groupes invertébrés (*odonates, rhopalocères, orthoptères, araignées, mollusques, écrevisses*)
- pour les autres données : validation manuelle uniquement
- une valeur de validation automatique :
  - peut être remplacée par une validation manuelle à tout moment
  - est recalculée si les données sont modifiées (*ex. cas de re-détermination*)





### 3 - Améliorer la qualité des données : validation scientifique

#### Critères utilisés pour la validation automatique :

- localisation :
  - à partir de l'habitat TAXREF (*FAUNA*)
  - à partir de la répartition connue de l'espèce : calcul entre répartition départementale, maillage et localisation de l'observation
- périodes d'observations favorables (*FAUNA*)
- difficulté de détermination du taxon
- validation du producteur
- sont exclus : taxons avec statut de protection (*OBV-NA*)

#### Taxons prioritaires à la validation manuelle :

- validation en interne pour le pôle flore-fonge-habitat
  - taxons problématiques et/ou susceptibles de faire l'objet d'erreurs de détermination
  - taxons protégés
  - observations pour lesquelles l'observateur lui-même a émis un doute
- validation externe pour la faune, effectuée par des partenaires ciblés (*taxons, localisations*)
  - logique expert centrée (invalidation dûes à la phénologie par ex., ou par espèce...)

Les niveaux de validation régionale sont ensuite consultables pour chaque donnée des observatoires.

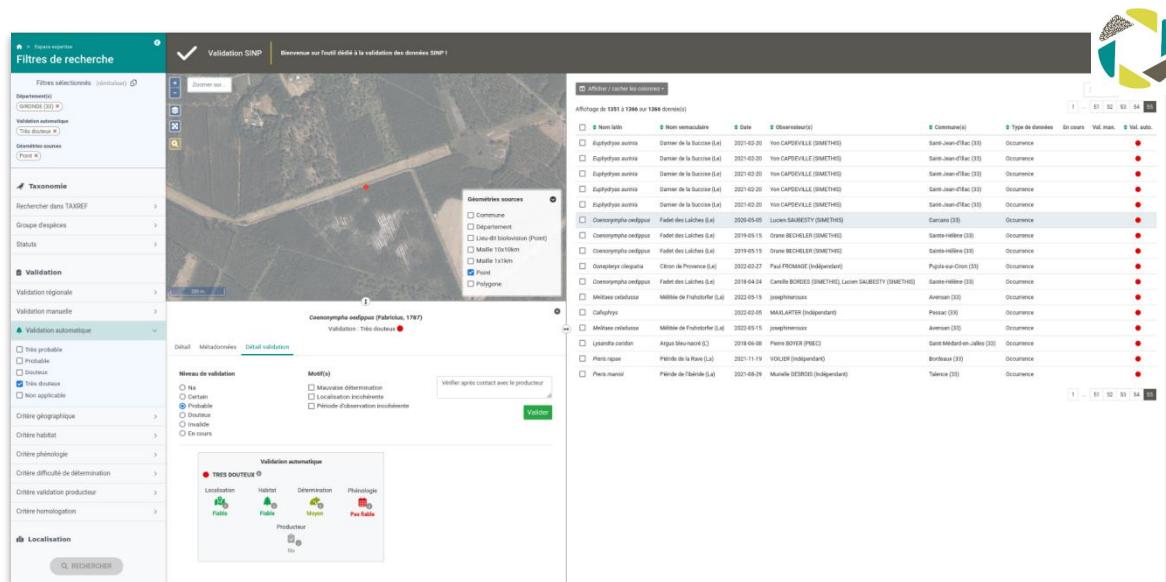
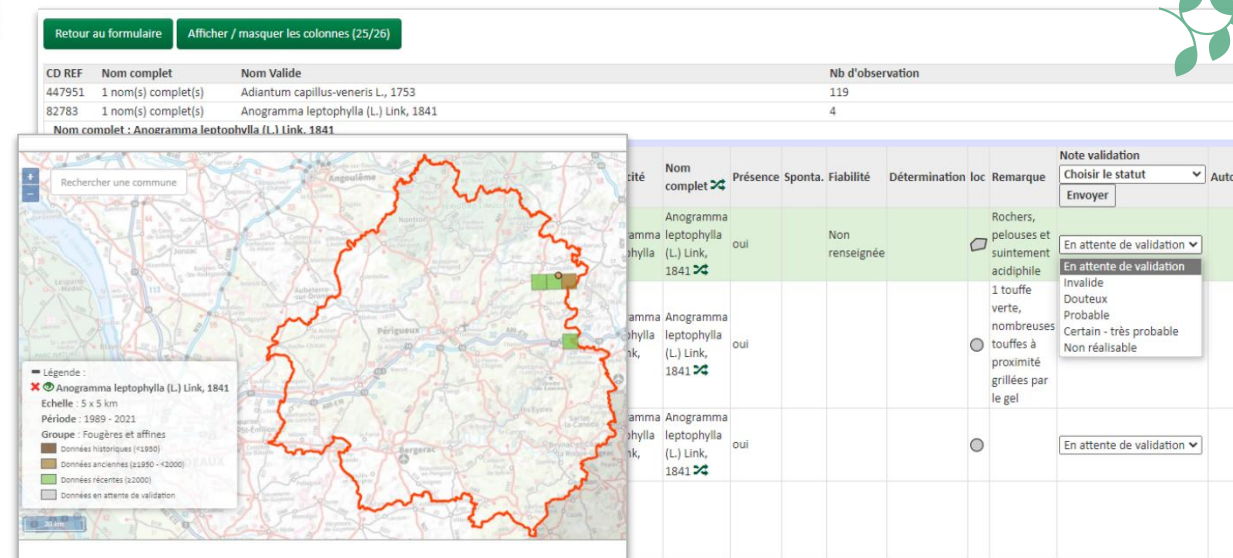




### 3 - Améliorer la qualité des données : validation scientifique

#### Module dédié à la validation manuelle :

- o accessible uniquement aux **personnes identifiées comme validateurs** (*droits gérés par département et groupe taxonomique*)
- o possibilité d'effectuer une **recherche par filtres** (*taxon, groupe taxonomique, note de validation, localisation...*)
- o validation possible pour une seule donnée ou par lots
- o affichage de différentes **informations apportant une aide à la validation** (*ex. commentaires liés à la détermination, note de fiabilité de la donnée, aire de répartition connue, espèces compagnes...*)

CD REF	Nom complet	Nom valide	Nb d'observation
447951	1 nom(s) complet(s)	Adiantum capillus-veneris L., 1753	119
82783	1 nom(s) complet(s)	Anogramma leptophylla (L.) Link, 1841	4

Nom complet	Présence Sponta.	Fiabilité	Détermination	Loc	Remarque	Note validation
Anogramma leptophylla (L.) Link, 1841	oui	Non renseignée			Rochers, pelouses et suintement acidophile	En attente de validation
Anogramma leptophylla (L.) Link, 1841	oui				1 touffe verte, nombreuses touffes à proximité grillées par le gel	En attente de validation
Anogramma leptophylla (L.) Link, 1841	oui					En attente de validation



**Améliorer la qualité  
des données en  
Nouvelle-Aquitaine :**

**4 - Précision et complétude**

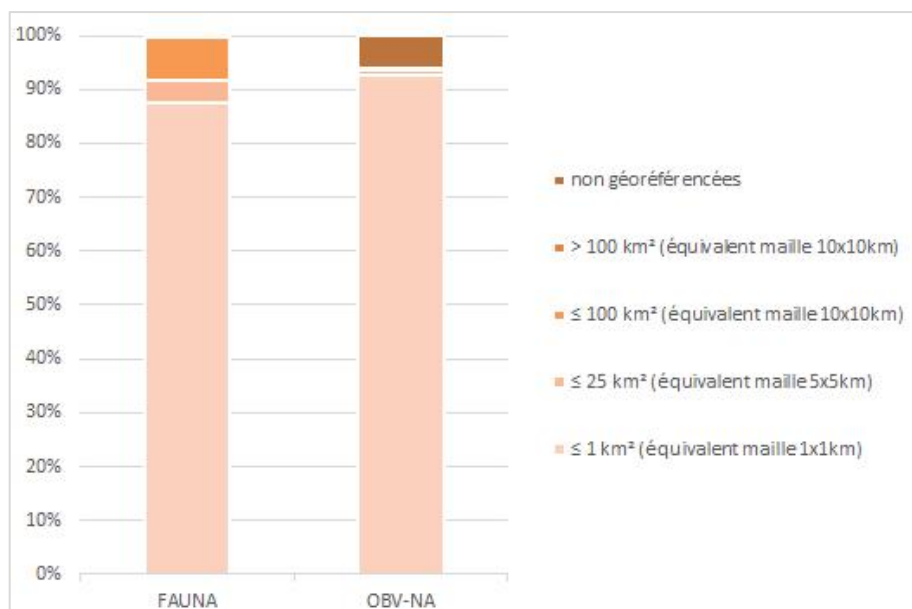




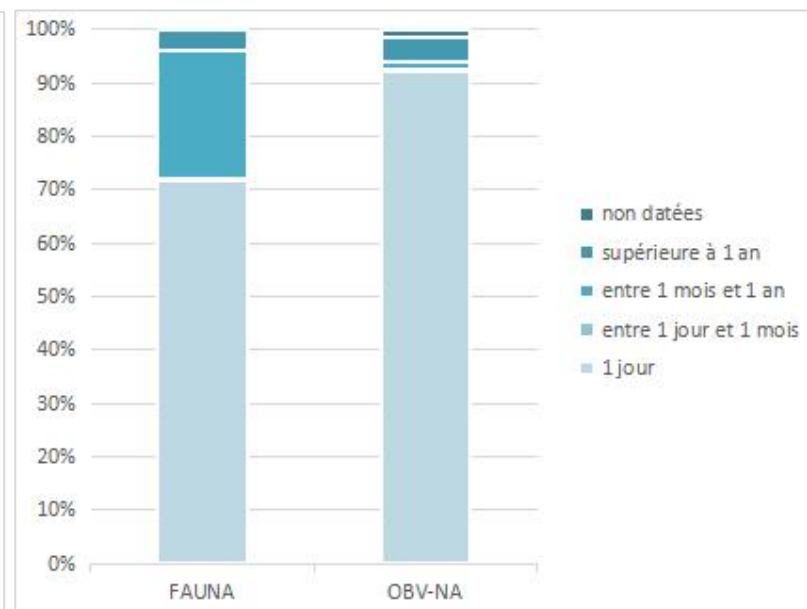
## 4 - Améliorer la qualité des données : précision et complétude

### Niveau de précision de la donnée :

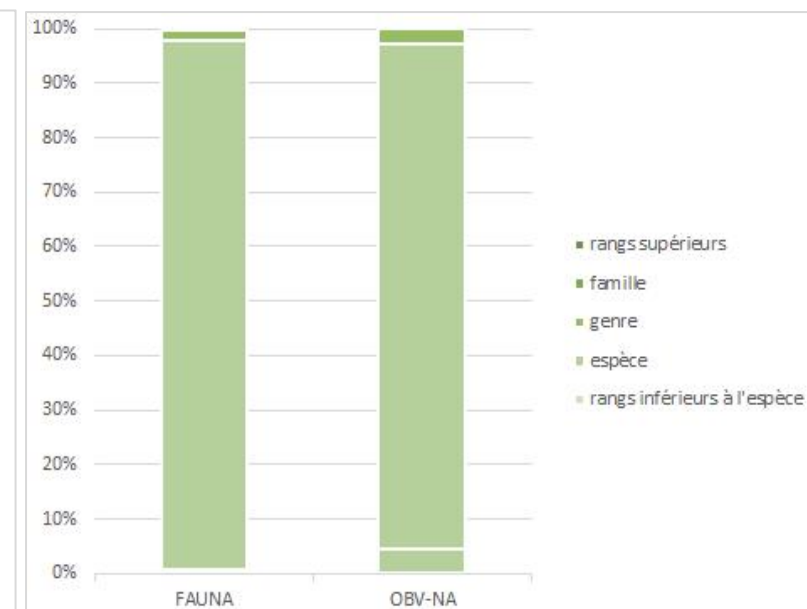
- Précision spatiale  
(localisation géographique)



- Précision temporelle  
(période d'observation)



- Précision taxonomique  
(détermination)







## 4 - Améliorer la qualité des données : **précision et complétude**

### Complétude des données et métadonnées :

Elle correspond au **niveau de renseignement des informations complémentaires** associées à une donnée/métadonnée

*métadonnées : description du protocole de collecte, indication de l'ensemble des acteurs...*

*observation : effectifs observés, comportements et stade de vie, spontanéité (cultivée, échouée...), présence/absence, phénologie, cohérence temporelle...*

#### ○ **Élément essentiel à la qualité** d'une donnée :

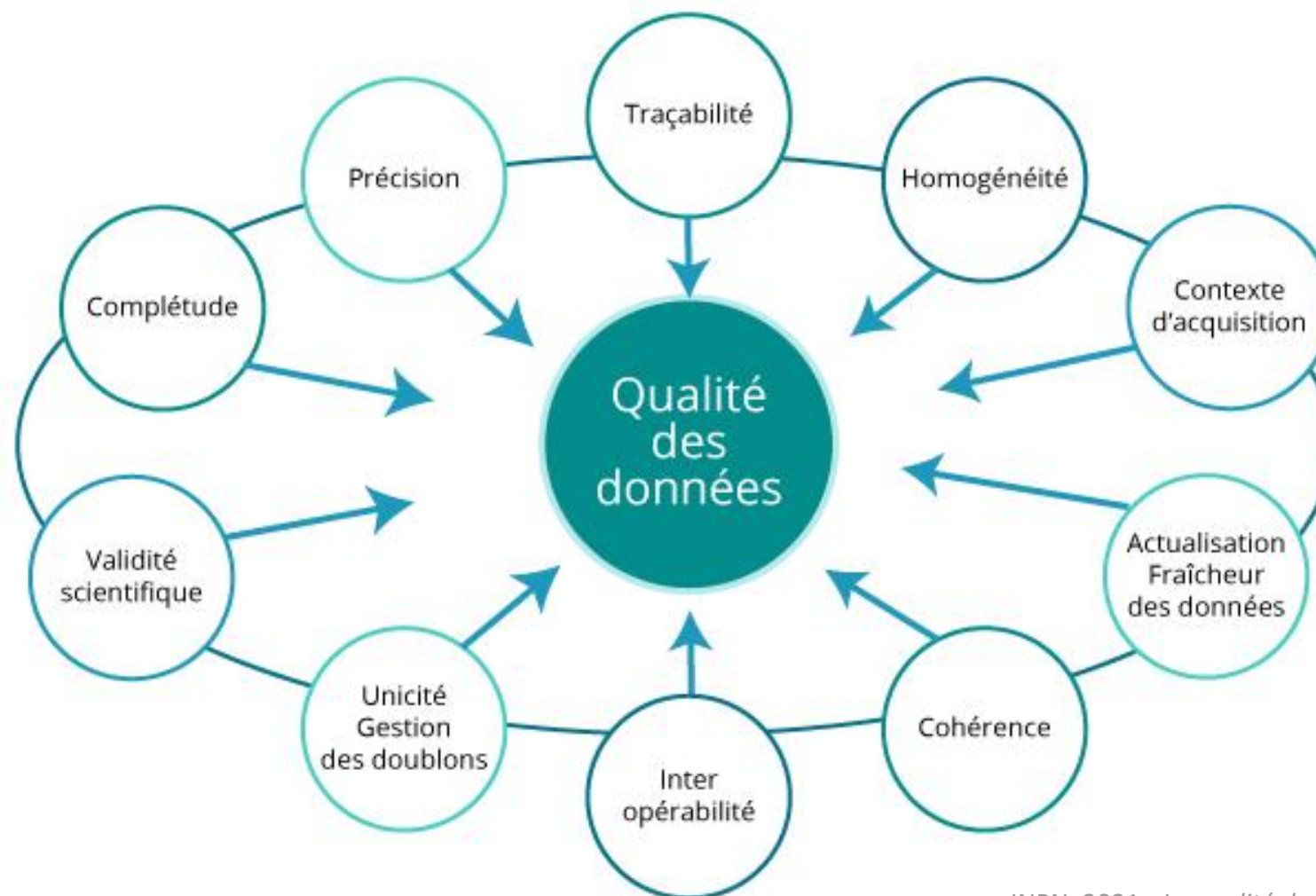
- pour l'**évaluation de la sensibilité** et du niveau de diffusion (*ex. reproduction pour la faune*)
- permet ou non la **validation régionale** (*preuve photo, note de fiabilité, statut de détermination...*)
- prise en compte lors de **programmes scientifiques** (*donnée opportuniste ou protocolée, type d'observation...*)

#### ○ A défaut, recherche d'informations parmi les champs non "standardisés" (*commentaires et remarques*)

- manuellement : étape très chronophage
- développement de méthodes automatisées : analyse automatique de commentaires selon une liste de motifs textuels pour compléter les données (reproduction, mortalité, fiabilité...)



## Améliorer la qualité des données : bilan



source : INPN, 2021 - *La qualité des données, un enjeu majeur pour le SINP*



Merci  
de votre attention