

Valoriser les données issues des Sciences Participatives: enjeux et défis méthodologiques

Pascal Monestiez,
BIOSP-INRA Avignon, CEBC-CNRS Chizé

Groupe de travail CiSStats

INRA-MIA, MNHN, CNRS-INEE, IRSTEA, ONCFS, AAMP
et en associatif LPO, TelaBotanica, APECS, Cybelle, CPIEs, ...

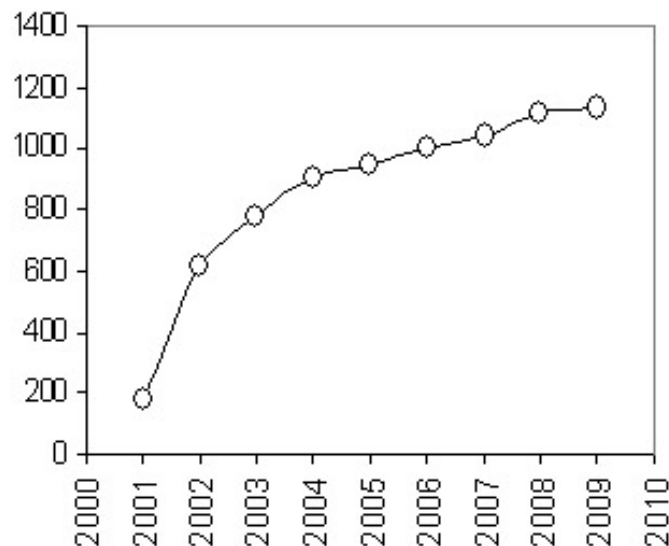
Développement des Sciences Participatives

- Un nombre de données en très forte croissance
- Le nombre de programmes de SP lui aussi en forte croissance : hétérogénéité des objectifs et des moyens, principe collaboratif
- Sources de données multiples, coexistence de protocoles différents : métadonnées faibles à absentes
- Très grand nombre de participants différents: intérêt et inconvénients du crowdsourcing

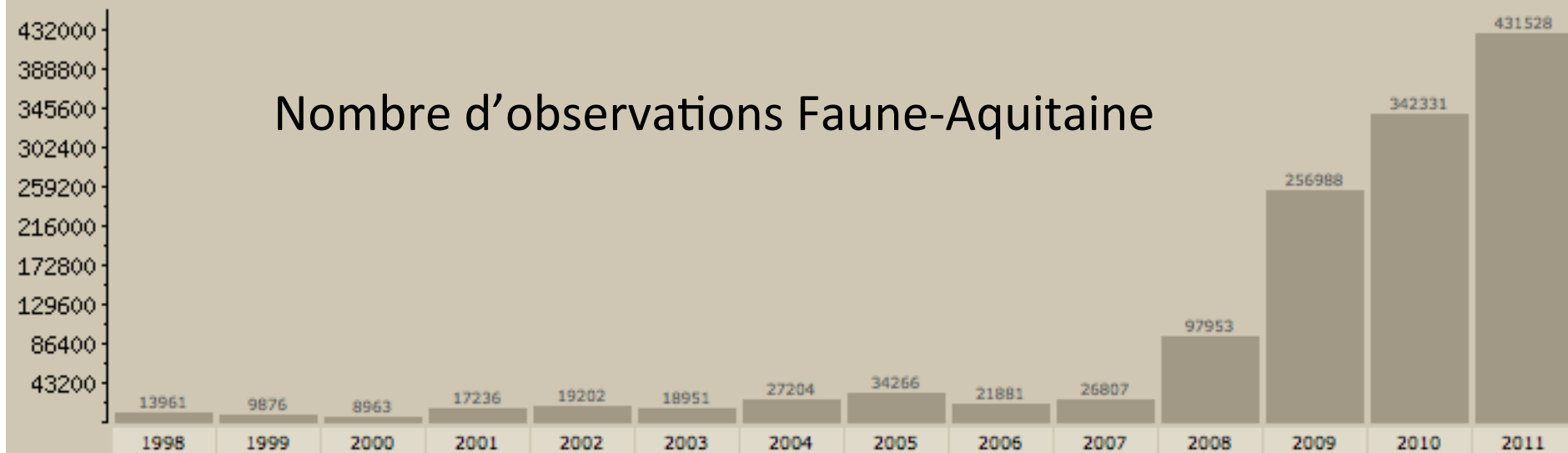
QUELQUES EXEMPLES

- Engouement pour les programmes Participatifs
- Changements technologiques
- Démarche Top-Down vs démarche Bottom-Up

Participants au programme STOC

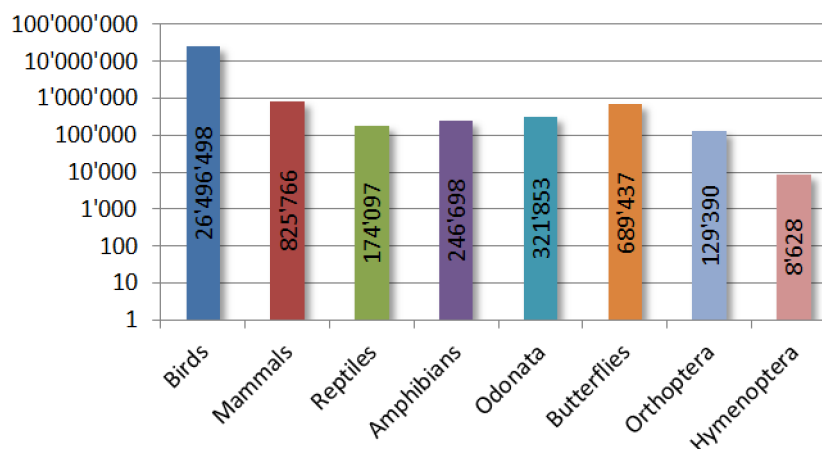


Nombre de contributions annuelles



Agrégation de données

- Biolovision : 11 ans d'expérience sur toute l'Europe, 70 Mobs, 26000 contrib. actifs,
- Pour la France



- GBIF (540 M observations, 1.6 M species, 740 data publishers),
- Ecoscope (metadonnées), Fishbase ...

Potentiel en suivi de Biodiversité

- Suivi global de la biodiversité (ensembles de taxons), suivi d'espèces cryptiques mais essentielles dans le fonct. des écosystèmes
- Possibilité de travailler à large échelle (région, Europe), et aussi de densifier avec un grain fin
- Travailler sur le court et long terme (détection de tendances, de ruptures, alertes)
- Substituts (ou compléments) à des systèmes institutionnels complexes et coûteux.

Données qui deviennent incontournables pour les scientifiques (écologie, conservation, changement climatique...)

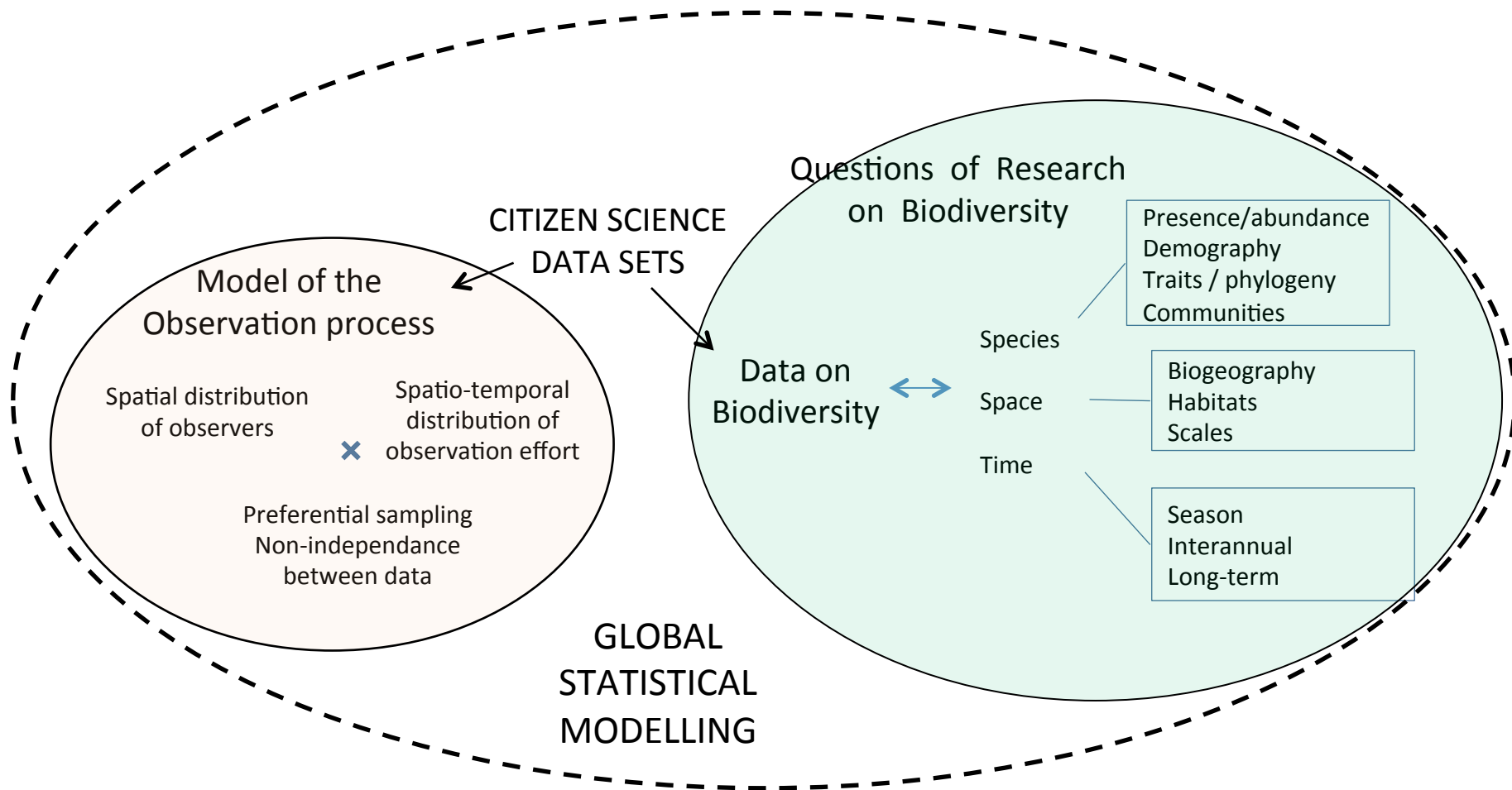
Hétérogénéité

- Caractéristique principale des données en SP :
 - **Hétérogénéité en répartition** (spatiale, temporelle) : le grand nombre « ne bouche pas les trous ». Hiver–été, périurbain–rural, milieu riche–milieu pauvre
 - **Hétérogénéité des observateurs** : bénévoles experts, amateurs débutants, scolaires, etc...
 - **Hétérogénéité des modes d'observations** : observations fortuites, observations dirigées, observations planifiées.
- A laquelle s'ajoute des **interactions complexes** (entre modes, entre observateurs, échantillonnage préférentiel) et différents niveaux de censure (espèces courantes vs espèces rares,)



Sources des principales difficultés méthodologiques

Un modèle générique global :



Modèle qui prend en compte

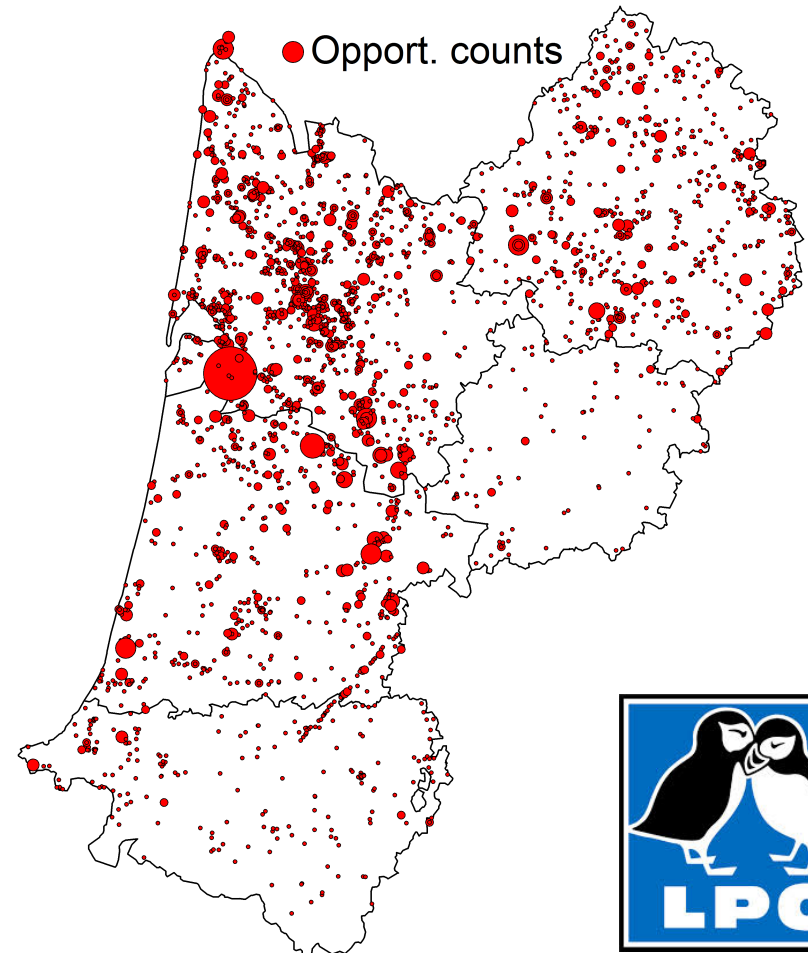
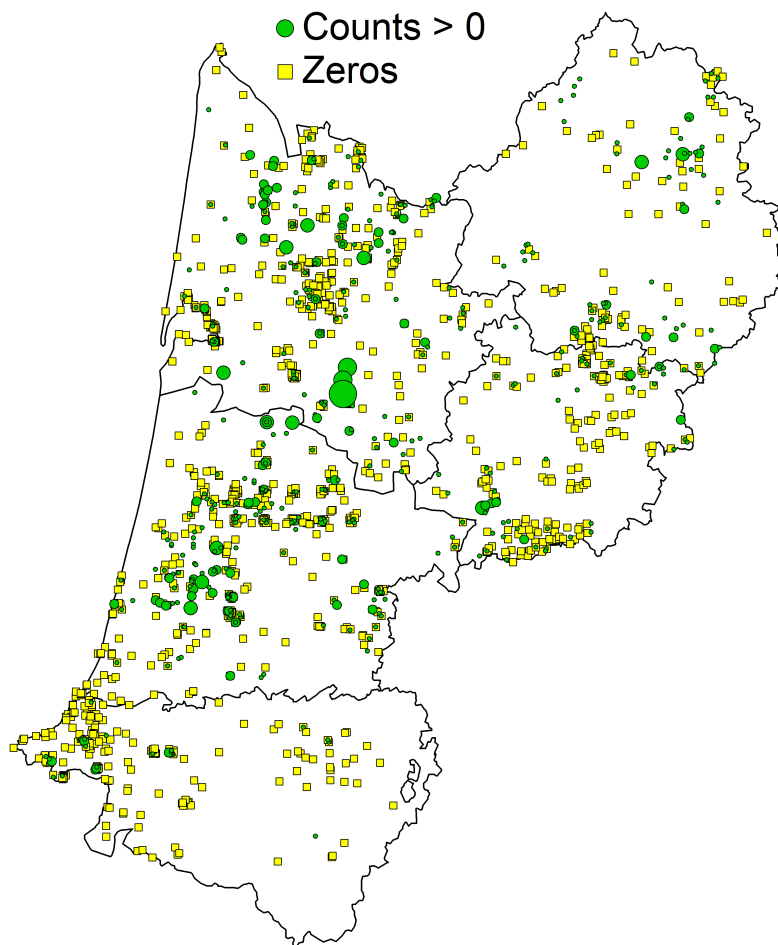
- l'hétérogénéité des observateurs, leur répartition, les taux d'erreur, les comportements.
- les censures : pourquoi ne fait-on pas l'effort de transmettre l'observation d'espèce courante, alors qu'une espèce rare fera l'objet d'observations redondantes ?
- l'échantillonnage préférentiel : pourquoi les observateurs bénévoles vont d'abord là où il ont le plus de chance de voir des espèces rares ?

Données LPO – Faune Aquitaine

un exemple simple : le Coucou



- Deux types de données :
type « formulaire » avec zéros et type « opportuniste »



Choix de modélisation

- Modéliser **le spatial** sous forme de champ de densité (effet aléatoire) cartographiable (l'effet environnement pouvant ensuite être ajouté en effet fixe).
- Combiner les données de différents types :
 - Avec **zéros** et efforts, et **opportunistes** pures
 - **Plurispécifique** (combinaison d'un petit nb. d'espèces)
- Prise en compte d'un effet **observateur** (effet aléatoire sur l'observateur : effort, efficacité), mais :
 - Pas d'erreur d'identification d'espèce (amateurs éclairés)
 - Pas d'interaction complexe
- **Cadre bayésien** pour bénéficier des approches **INLA** (rendre l'estimation acceptable en temps et moyens)

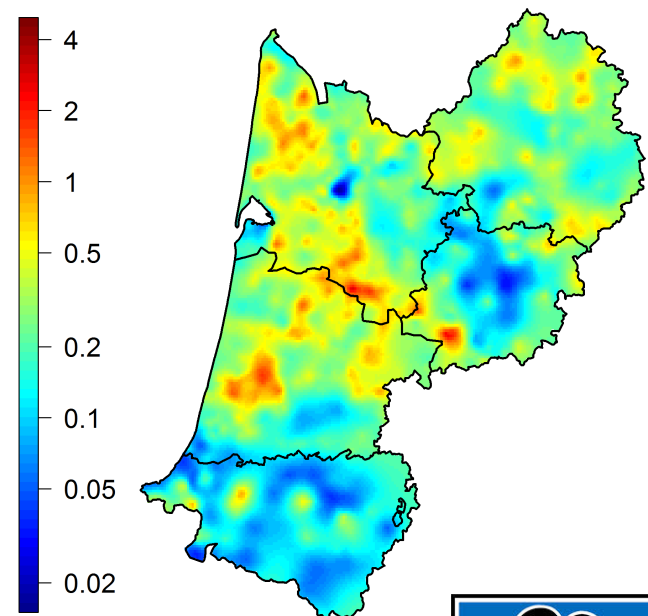
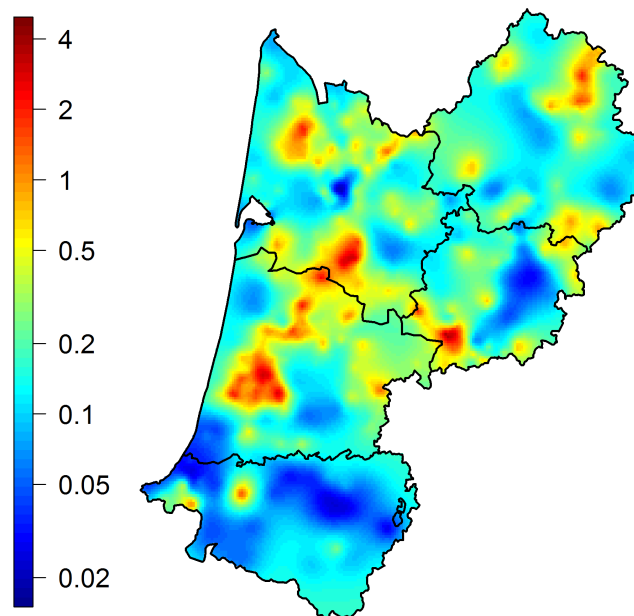
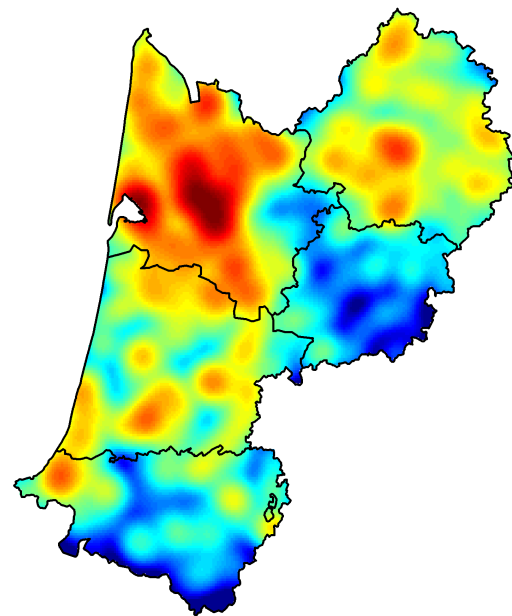
Données observations LPO – Faune Aquitaine

Résultats : exemple du Coucou gris

Opportunistes seules
estimation par kernel

« formulaires » seules
sans les observateurs

2 jeux de données +
effets observateur



Défis méthodologiques

- Le **nombre de paramètres** augmente avec le nombre d'espèces, le nombre d'observateurs différents, le nombre de sources et de protocoles différents qu'il faut agréger.
- Modèles statistiques de plus en plus complexes : **nombre croissant d'interactions** que l'on ne peut pas négliger
- Développement de modèles (construction de modèles hiérarchiques à plusieurs niveaux latents), des **méthodes d'inférence en grande dimension**, des méthodes numériques et algorithmiques pour de telles inférences.

Premières conclusions

- Le grand nombre ne corrige pas les biais et ni les problèmes d'échantillonnage préférentiel
- Corriger ces biais demande de rester très près des producteurs de donnée. Connaissance des processus d'observation et des observateurs (impossible à faire une fois les données agrégées)
- Le grand nombre permet de modéliser les interactions si il y a suffisamment d'observations par observateur.
- Favoriser la fidélisation des observateurs (profils), le pluri spécifique (corrections croisées), gérer le compromis entre le grand nombre et la complexité du protocole.
- Si possible, croiser les données SP avec des données de « référence » : extension spatiale et temporelle des dispositifs institutionnels.

Merci de votre attention

Groupe de travail CiSStats

INRA-MIA, MNHN, CNRS-INEE, IRSTEA, ONCFS, AAMP
et en associatif LPO, TelaBotanica, APECS, Cybelle, CPIEs, ...