Proposition d„un protocole de validation de donnéespour le GRETIA

Loïc CHÉREAU[1] et Philippe ZORGATI[2]

Mots-clés – Méthodologie, base de données, euarthropodes continentaux, identification.

Résumé – Les auteurs proposent une méthodologie de validation de données entomologiques, élaborée dans le cadre de la dynamique du groupe « Carabiques » du GRETIA.

Abstract – The authors propose a methodology for the validation of entomological data, elaborated within the framework of the dynamics of the "Carabidae" group of GRETIA.

Contexte

Les inventaires naturalistes connaissent depuis quelques années un essor particulier. Les entomologistes prennent de plus en plus conscience de l'intérêt de leurs « données » et de l'utilité de les mettre à disposition au sein de bases régionales ou nationales. Les techniques liées à Internet offrent également aux entomologistes, qu'ils soient débutants ou confirmés, des outils nouveaux destinés à rassembler leurs témoignages (forums, bases de données en ligne). Se pose alors la question de la validation de ces données compilées.

Telle donnée est-elle fiable ? Le taxon « X » n'a-t-il pas été confondu avec le taxon proche « Y » ? Le signalement de la capture de l'espèce « X » à tel endroit est-il plausible ? Que fait-on en cas de doute ? Comment ne pas préjuger de l'expérience de l'auteur dans le processus de validation ? Soulignons qu'une donnée fausse ne résulte pas nécessairement d'une erreur d'identification. Une erreur d'étiquetage ou une faute dactylographique peuvent aussi en être à l'origine.

Depuis 2013, nous avons tenté de répondre à ces questions. Amateurs de carabiques, la base de données du GRETIA nous a offert une matière intéressante. Elle regroupe à la fois des citations bibliographiques anciennes et des données d'observateurs contemporains. Pour prendre en charge la validation des données armoricaines de carabiques « Carabidae Latreille, 1802 » (LOBL & SMETANA, 2003), nos réflexions nous ont amenés à concevoir et expérimenter une méthodologie qui nous a permis de valider (ou non) une 1ère série de 4 500 données. Elle n'a pas la prétention d'ôter toute subjectivité à la démarche mais de proposer au moins une méthode reproductible d'un échantillon de données à l'autre, permettant d'argumenter sans ambiguïté le processus de validation.

La perspective que ce protocole de validation puisse être largement utilisé au GRETIA a également motivé notre travail.

Figure 1. Réunion du groupe « Carabiques » autour du protocole de validation, Assemblée générale d'automne 2015 du GRETIA (Cliché : C. Mouquet).

Principes de validation

Il est bien évident qu'il est impossible de valider « de visu » tous les exemplaires faisant l'objet d'un signalement. Deux options s'offrent à nous : refuser ceux que nous n'aurions pas vus ; les accepter sous certaines conditions, de façon à minimiser le risque de « passer à côté » d'erreurs ou d'omissions.

A ce sujet, nous avons démarré nos réflexions par la lecture des principes de validation utilisés dans le « Catalogue des Coléoptères de Rhône-Alpes : Coléoptères Carabiques et Cicindèles» (COULON J. *et al.*, 2000) que nous résumons ici :

[1] Les Fresnes, F-50680 Couvains, <loic-chereau@wanadoo.fr>

[2] 5, rue de la Jouennerie, F-50130 Cherbourg-Octeville, <p.zorgati@orange.fr>

1er cas *- L„espèce citée est rare et/ou facile à confondre avec un autre taxon intéressant et/ou très peu citée et/ou citée d„un secteur géographique a priori insolite : la consultation du spécimen est alors demandée.*

2e cas *- L„espèce citée est commune et largement répandue : la consultation n„est pas nécessaire. Deux cas de figure se présentent alors : la bête est bien identifiée et c„est tant mieux (mais on ne le saura pas) ; la bête est mal identifiée et dans ce cas :*

- l'espèce identifiée sous le nom X est à rattacher au taxon Y tout aussi banal : l'erreur sera noyée dans la masse et ne saurait remettre en cause la répartition et a fortiori la présence des 2 taxons concernés.
- l'espèce est à rattacher à un taxon intéressant. La citation de ce taxon sera alors malheureusement occultée.

Nous avons également étudié le cadre méthodologique utilisé pour la validation des données de rhopalocères pour l'INPN (Dupont, 2014). Cependant, certains critères présupposent un niveau de connaissance de l'écologie ou de la répartition des espèces, tel l'habitat du taxon ou encore sa phénologie, qui rend la méthode de validation non transposable à d'autres groupes biologiques moins connus, tels les Carabidae. En outre, si la méthodologie de validation que Dupont propose est relativement détaillée, certains critères ne sont pas clairement définis, comme par exemple : « la détermination a été réalisée avec des éléments associés suffisants » ou encore « la confusion avec un autre taxon dans le département est statistiquement réduite ». La possibilité d'une double identification du taxon est en outre peu mise en avant, limitée au cas de l'examen de données douteuses, et l'écueil d'un distinguo entre spécialiste et non-spécialiste reste peu nuancée offrant une porte ouverte à la controverse.

Démarche

Groupes d„espèces

La 1ère tâche a consisté à dresser une liste des espèces certaines, probables ou possibles de la dition et à l'intérieur de celle-ci, celles qui nous paraissaient présenter un intérêt particulier à être examinées, soit en raison de leur « rareté », soit en raison de leur difficulté de détermination. La compilation de catalogues régionaux anciens avec le récent catalogue des coléoptères de France (TRONQUET coord., 2014), enrichies des données de nos propres collections entomologiques, a constitué une solide base pour réaliser ce travail.

Une espèce commune en vallée de la Loire peut être improbable en vallée de la Touques. Le choix de l'échelle de travail peut donc fortement influencer la répartition des espèces de carabiques dans tel ou tel groupe d'espèces. Dans l'état actuel de nos connaissances et conscients de ne pouvoir utiliser ce protocole sur l'ensemble du territoire du GRETIA (pour le moment !), la ventilation des espèces par groupe est envisagée uniquement pour la Normandie occidentale (Calvados, Manche et Orne).

Nous distinguons trois groupes :

Groupe A : espèces communes et/ou répandues et d'identification estimée facile par le groupe des amateurs de carabiques du GRETIA.

Groupe B : espèces que le validateur souhaiterait voir :

- soit parce qu'elles sont « rares » ;
- soit parce qu'elles sont de détermination délicate et qu'on pourrait passer à côté d'une « bonne bête » : cette démarche est peut-être ambitieuse, mais elle permet de réduire les cas de taxons intéressants passant inaperçus car confondus avec des espèces communes et répandues.

Groupe C : espèces douteuses (en dehors de leur biotope, très éloignées de leur aire de répartition connue), c'est-à-dire que le groupe des amateurs de carabiques du GRETIA juge fortement ou totalement improbables sur la dition. Dans le cas des carabiques, sont notamment ainsi catégorisés la totalité des *Trechinae* cavernicoles, les espèces endémiques alpines ou pyrénéennes ainsi que les espèces strictement méditerranéennes.

Processus de validation : résultats et cas.

Trois résultats sont envisagés pour les données ayant suivi le processus de validation :

- données validées ;
- données en cours de validation ;
- données non validées.

Pour chacun de ces résultats, une diversité de situations existe. Par exemple, pour les données validées, nous aurions pu utiliser des qualificatifs tels que « certain », « douteux », « probable ». Mais il aurait fallu introduire une notion de taux de probabilité par trop artificielle (Très probable ? Peu probable ? Presque certain ?). Il nous a semblé plus judicieux d'introduire une notion de « **cas de validation** ». Chaque cas de validation correspond à un cheminement intellectuel, qui intègre principalement :

- le groupe d'espèces (A, B ou C) auquel a été rattaché le taxon considéré,
- le fait que le validateur ait vu ou pas le spécimen,
- le fait que l'identification du spécimen soit simple ou double (et réalisée de façon indépendante),
- le fait que la donnée ait été publiée ou non (ce caractère n'est pas retenu comme suffisant en soi).

Nous considérons ici comme « publications » l'ensemble des « productions écrites diffusées », qu'il s'agisse de rapport d'études, d'articles de revues ou d'ouvrages. Il est supposé qu'à partir du moment où une donnée a fait l'objet d'une diffusion écrite, la détermination a été faite avec soin. Derrière cette généralité, gardons présent à l'esprit que demeure le risque de données erronées et tout de même publiées.

Différents cas ont donc été déterminés°:

Donnée validée :

- **cas 1** : détermination du spécimen par au moins deux personnes, dont le validateur
- **cas 2** : seul le validateur a identifié correctement le spécimen
- **cas 3** : donnée publiée pour une espèce du groupe A
- **cas 4** : donnée publiée pour une espèce du groupe B pour laquelle :

  aucun changement taxonomique depuis la date de publication

  et

  aucun risque d'erreur d'identification n'est avéré avec les ouvrages de référence de l'époque, notamment lorsque ces publications contiennent des figures permettant de confirmer l'identité de l'espèce sans aucune ambiguïté

  et

  le contexte écologique est cohérent avec les éléments d'autoécologie connus pour l'espèce »

- **cas 5** : donnée non publiée, espèce du groupe A.

Donnée non validée :

- **cas a** : Donnée publiée pour une espèce du groupe B et pour laquelle on signale une « évolution taxonomique (révision du complexe d'espèces par exemple) depuis la date de publication de la donnée et/ou un risque d'erreur d'identification avéré avec les ouvrages de référence de l'époque et/ou contexte écologique incohérent avec les éléments d'autoécologie connus pour l'espèce » ;
- **cas b** : Donnée publiée pour une espèce du groupe C et pour laquelle est présumée l'impossibilité d'avoir accès au spécimen ;
- **cas c** : Donnée non publiée pour une espèce du groupe B ou C et pour laquelle est présumée l'impossibilité d'avoir accès au spécimen.

La numérotation des cas ne correspond pas à un degré de certitude hiérarchisé.

La non-validation ne signifie pas l'invalidation. Les données saisies avant validation dans la base de données du GRETIA se voient attribuées par défaut le statut « à valider ». Les données extraites de la base de données du GRETIA pour être étudiées par le prisme du protocole de validation sont qualifiés « actuellement soumises au validateur ».

En résumé

La validation, telle que nous l'envisageons, s'exerce à partir :

- d'une exportation de la base de données du GRETIA à une date considérée, pour un groupe considéré ;

- de l'usage d'une clé de validation et d'une liste d'espèces associées ventilées en 3 groupes, par une personne à une date donnée.

Ainsi, du processus de validation résultera, pour une donnée : le statut de validation, la date de l'opération de validation, le nom du validateur, la version de la clé de **validation utilisée** (dont nous proposons ici une version 1), la version de la classification par groupe des espèces et le cas échéant, toute précision utile du validateur sous forme de remarque.

Préalablement, une suppression des doublons doit être réalisée manuellement.

Le processus de validation mériterait d'être automatisé pour partie, d'une part pour minimiser les risques d'erreurs dues à la manipulation des données et, d'autre part, pour économiser le temps du validateur.

Après avoir validé les données, l'édition des cartes et de leurs diagrammes phénologiques de chaque espèce permettra un dernier « passage en revue ». Il pourra être réalisé collectivement, par le groupe des amateurs de carabiques du GRETIA. De façon argumentée, le statut de validation de certaines données pourra ainsi être modifié car le processus de validation proposé ici sera sans nul doute amené à être affiné dans le souci constant d'améliorer nos méthodologies de travail.

Conclusion

En présentant cette méthodologie de validation de données naturalistes, nous n'avons évidemment pas la prétention de proposer une démarche exempte de toute erreur, approximation ou omission.

Les trois objectifs qui nous ont guidés ont été :

- de proposer une démarche reproductible d'un échantillon à l'autre et d'un validateur à l'autre ;

- de clarifier les processus de validations sur des critères aussi objectifs que possible ;

- d'expliquer les non validations.

Nous avons conçu cette clé de validation pour l'appliquer au cas des Carabiques, groupe nécessitant souvent un important effort de détermination. Il est sans doute possible d'y trouver des simplifications applicables à des groupes taxonomiques plus abordables.

L'absence d'une « alerte phénologique » dans le processus de validation peut être regrettée. Ce sera sans nul doute un point à améliorer notamment grâce aux évolutions en matière de programmation informatique.

L'évaluation du niveau d'expertise du validateur demeure un sujet délicat. Le fait d'avoir étudié les spécimens types des taxons en question est généralement le signe d'un haut niveau d'expertise. La capacité à travailler en groupe de coléoptéristes au sein du GRETIA et plus largement avec la communauté entomologique est un gage de qualité dans le processus de validation des carabiques. Car le travail de validation est une chose trop sérieuse pour n'être confié qu'à un seul entomologiste ! Nous pouvons témoigner comme les interactions entre débutants et spécialistes sont toujours constructives pour les uns comme les autres dans leurs cheminements naturalistes et scientifiques respectifs.

Il y tant à découvrir que la taxonomie est en perpétuelle évolution. Insistons donc pour chaque témoignage d'espèce sur l'importance de disposer d'un spécimen de référence déposé dans une collection publique qui seul permettra son examen par les futures générations d'entomologistes.

Remerciements.- Nous adressons tous nos remerciements aux personnes ayant porté leur regard critique sur ce travail pour nous permettre de le faire évoluer : Rémy Ancellin, Eric Drouet, François Dusoulier, Jean-Pierre Favretto, Etienne Iorio, Mathieu Lagarde, Adrien Simon ainsi que les membres du groupe « Carabiques » du GRETIA.

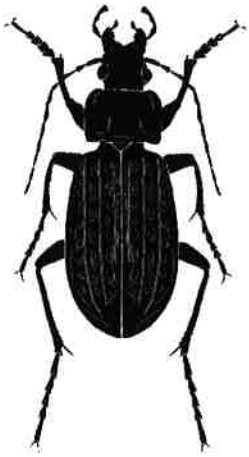
Bibliographie

COULON J. *et al.*, 2000.- Coléoptères de Rhône-Alpes : Carabiques et Cicindèles, Muséum d'Histoire naturelle de Lyon et Société linnéenne de Lyon pp 18-19

DUPONT P. 2014.- Cadre méthodologique de l'inventaire national des Rhopalocères et Zygènes de France métropolitaine. Partie I. Muséum National d'Histoire Naturelle, Paris. Rapport SPN 2014 - 23. 28 pp

LOBL I. & SMETANA A. 2003.- Catalogue of Palaearctic Coleoptera, vol 1 : Archostemata, Myxophaga, Adephaga. Apollo Books, Stenstrup. 819 pp

TRONQUET M., Coord. (2014).- Catalogue des Coléoptères de France. Supplément au tome XXIII de la revue R.A.R.E., 1052 p.



Clé récapitulative

groupe A : espèces communes et/ou répandues et d'identification estimée facile par le validateur

groupe B : espèces que le validateur souhaiterait voir (demande motivée)

groupe C : espèces douteuses (en dehors de leur aire de répartition connue dans la région considérée)

1 (8)	**Validateur disposant du spécimen**	
2 (3)	Spécimen identifié préalablement **et** confirmation de l'identification par au moins 1 validateur (double détermination)	**Validation (cas 1)**
3 (2)	Spécimen confié non identifié préalablement **ou** remise en cause de l'identification	
4 (5)	• Validateur incapable d'identifier le spécimen (mauvais état du spécimen, taxon dont l'un des deux sexes n'est pas identifiable spécifiquement, limite de la compétence du validateur,...)	**Validation en cours** et référencement dans la BDD de la personne dans la collection de laquelle sera conservé le spécimen
5 (4)	• Validateur capable d'identifier le spécimen	
6 (7)	□ Nouvelle identification réalisée par 1 validateur (donnée mise à jour)	**Validation (cas 2)**
7 (6)	□ Nouvelle identification réalisée par 1 validateur **et** par au moins une autre personne (double détermination) (donnée mise à jour)	**Validation (cas 1)**
8 (1)	**Validateur ne disposant pas du spécimen**	
9 (18)	Donnée publiée	
10 (11)	• Espèce du groupe A	**Validation (cas 3)**
11 (10)	• Espèce du groupe B ou C	
12 (15)	• Espèce du groupe B	
13 (14)	□ Pas de changement taxonomique depuis la date de publication **et** pas de risque d'erreur d'identification avéré avec les ouvrages de référence de l'époque **et** contexte écologique cohérent avec les éléments d'autoécologie connus pour l'espèce	**Validation (cas 4)**
14 (13)	□ Evolution taxonomique (révision complexe d'espèces par exemple) depuis la date de publication de la donnée **et/ou** risque d'erreur d'identification avéré avec les ouvrages de référence de l'époque **et/ou** contexte écologique incohérent avec les éléments d'autoécologie connus pour l'espèce	**Non validation** et blocage de la donnée **(cas a)**
15 (12)	• Espèce du groupe C	
16 (17)	□ Possibilité présumée d'accéder au spécimen (observateur contemporain, ...)	**Validation en cours** et demande à l'observateur du spécimen (une fois le spécimen obtenu, recommencer la clé au début / si pas d'envoi du spécimen après 6 mois : non validation **(cas b)**
17 (16)	□ Impossibilité présumée d'avoir accès au spécimen	**Non validation** et blocage de la donnée **(cas b)**
18 (9)	Donnée non publiée	
19 (20)	• Espèce du groupe A	**Validation (cas 5)**
20 (19)	• Espèce des groupes B **ou** C	
21 (22)	□ Possibilité présumée d'accéder au spécimen (observateur contemporain, ...)	**Validation en cours** et demande à l'observateur du spécimen (une fois le spécimen obtenu, recommencer la clé au début / si pas d'envoi du spécimen après 6 mois : non validation **(cas c)**
22 (21)	□ Impossibilité présumée d'avoir accès au spécimen	**Non validation** et blocage de la donnée **(cas c)**