Sciences citoyennes et qualité des données sur la biodiversité : un faux problème ?

Bastien CASTAGNEYROL
INRAE - UMR BIOGECO

Christophe BOTELLA
CNRS - LECA

Benoît FONTAINE
MNHN - CESCO

Quelle confiance accorder aux données issues des programmes de sciences et recherches participatives (SRP) si elles sont collectées massivement (crowdsourcing) par des volontaires dont le niveau d'expertise est variable et inconnu ? Cette question en appelle immédiatement une seconde : quel crédit accorder aux résultats scientifiques fondés sur ces données ? À partir d'un examen de la littérature scientifique internationale et de cas d'études de projets de SRP menés sur le territoire français dans le domaine de l'écologie, nous discutons ces interrogations et proposons des éléments de réponse. Nous montrons que la fiabilité des données issues de crowdsourcing est entièrement dépendante des questions auxquelles elles répondent, et que de nombreux points de contrôle existent — en amont et en aval des observations — pour assurer la fiabilité des inférences réalisées à partir de ces données. Nous concluons que, pour peu que le protocole de collecte des données soit en adéquation avec la question scientifique posée, la fiabilité des données collectées en masse par le grand public n'est pas problématique, et que les limitations intrinsèques des SRP fondées sur le crowdsourcing sont largement compensées par les opportunités qu'elles offrent aux niveaux scientifique et sociétal.

Faut-il avoir un doctorat pour contribuer à la production de connaissances scientifiques ? Le développement massif des programmes de sciences et recherches participatives (SRP) dans lesquels chercheurs professionnels et volontaires[1] collaborent étroitement à l'acquisition, la validation, l'analyse et l'interprétation de données scientifiques suggère que non. Toutefois, ce constat ne remet pas en question le haut niveau d'expertise nécessaire à la production des connaissances scientifiques. C'est plutôt une invitation à s'interroger sur les différentes étapes de la production des données qui garantissent leur fiabilité et leur reproductibilité. Dans cet article, nous nous intéressons aux SRP se limitant à l'implication du public dans la collecte de données à grande échelle (crowdsourcing) dans le domaine de l'écologie. Après une présentation rapide de la diversité de ces programmes illustrée par quelques exemples emblématiques, nous posons une question cruciale : les données collectées en masse par des volontaires au degré d'expertise variable sont-elles de qualité suffisante pour contribuer à l'avancée des connaissances scientifiques ? Nous y répondons en

1 Les termes utilisés pour faire référence aux personnes participants aux programmes de SRP sont extrêmement variés (amateurs, citoyens, non-chercheurs, non-scientifiques professionnels, « volonpairs »…). Le choix des mots n'est pas neutre et véhicule implicitement un ensemble de représentations (Eitzel *et al.* 2017). Dans cet article, nous nous focalisons sur les programmes de sciences citoyennes se limitant à la collecte de données ou de matériel par le public, sans distinction d'âge, d'expertise, de citoyenneté. Nous qualifierons les participants de « volontaires », et de « scientifiques » ou de « chercheurs » les professionnels à l'origine de ces programmes, étant entendu que ces deux catégories peuvent être partiellement chevauchantes.

confrontant retours d'expériences et évaluations formelles publiées dans la littérature scientifique. Enfin, nous discutons les moyens qui peuvent être mis en œuvre pour garantir la qualité des données.

Pourquoi et comment faire appel aux volontaires pour collecter des données en masse ?

De quoi parle-t-on ?

Les collections naturalistes des muséums d'histoire naturelle doivent beaucoup aux naturalistes amateurs de sorte que les SRP ne sont pas une invention du XXIe siècle (voir l'article de Volny Fages dans ce numéro). Un des plus anciens programmes de SRP date du début du XXe siècle : le programme Christmas Birds Count invite les américains à répertorier la diversité des oiseaux, tous les ans depuis 1900, pendant la période de Noël[2]. Mais l'essor des SRP, et notamment de la collecte de données en masse, est très largement favorisé par le développement des technologies informatiques et des outils connectés[3]. Aujourd'hui, les SRP s'adressent à un public plus ou moins qualifié, pouvant participer depuis son canapé, son jardin, son quartier… avec un minimum de matériel (**Tableau 1**). Par exemple, le programme Penguin Watch propose aux volontaires de participer à l'étude de la dynamique des populations d'oiseaux marins à partir du dénombrement de pingouins sur des images issues de pièges photographiques[4]. L'observatoire de la biodiversité des jardins invite les volontaires novices à observer les papillons, bourdons et escargots dans leur environnement proche, contribuant ainsi à la compréhension de l'impact des sociétés humaines sur la biodiversité commune[5,6]. En parallèle de ces projets "tout public", existent également des programmes de SRP impliquant des naturalistes hautement qualifiés. C'est le cas du Suivi Temporel des Oiseaux Commun (STOC) qui s'adresse aux ornithologistes confirmés, depuis 1989[7]. Les algorithmes permettant l'identification automatisée des plantes photographiées au travers de l'application Pl@ntNet sont également alimentés par les données collectées par le grand public, et validées collaborativement par des botanistes confirmés.
Les projets de SRP fondés sur la collecte massive de données auxquels nous nous intéressons ici ont en commun d'avoir pour objectif premier la construction de nouvelles connaissances sur l'environnement ou la biodiversité ; à cet objectif scientifique peut s'ajouter un objectif social pouvant se décliner au niveau de l'individu (e.g., gain en compétences, reconnaissance sociale) et de la société dans son ensemble à travers la contribution des SRP à l'élaboration de politiques publiques[8].

Gagnant-gagnant ?

Les projets de SRP sont un apport inestimable à la recherche dans les domaines de l'écologie et de l'environnement. C'est particulièrement vrai dans le cadre de l'étude de la dynamique de la biodiversité dans le contexte des changements globaux[9], des effets du réchauffement climatique sur la structure des communautés et le fonctionnement des écosystèmes[10], de la surveillance de l'expansion des espèces exotiques envahissantes[11], de l'efficacité des politiques de conservation[12] ou encore de l'amélioration des plans de conservation des espèces menacées[13]. La raison en est simple : la participation de volontaires à la collecte de données scientifiques démultiplie les capacités

2 Silvertown J. (2009) A new dawn for citizen science. Trends in Ecology & Evolution 24(9):467–471.

3 Graham E., S. Henderson and A. Schloss (2011) Using mobile phones to engage citizen scientists in research. EOS Transactions 92(38).

4 Jones F. M., C. Allen, C. Arteta, J. Arthur, C. Black, L. M. Emmerson, R. Freeman, G. Hines, C. J. Lintott, Z. Macháčková, G. Miller, R. Simpson, C. Southwell, H. R. Torsey, A. Zisserman and T. Hart (2018) Time-lapse imagery and volunteer classifications from the Zooniverse Penguin Watch project. Scientific Data 5:180124.

5 Bergerot B., B. Fontaine, M. Renard, A. Cadi and R. Julliard (2010) Preferences for exotic flowers do not promote urban life in butterflies. Landscape and Urban Planning 96(2):98–107.

6 Muratet A. and B. Fontaine (2015) Contrasting impacts of pesticides on butterflies and bumblebees in private gardens in France. Biological Conservation 182:148–154.

7 Julliard R., F. Jiguet and C. Denis (2004) Common birds facing global changes: What makes a species at risk? Global Change Biology 10(1)148–154.

8 Adler F. R., A. M. Green and Ç. H. Şekercioğlu (2020) Citizen science in ecology: a place for humans in nature. Annals of the New York Academy of Sciences 1469:52–64.

9 Schmeller D. S., P.-Y. Henry, R. Julliard, B. Gruber, J. Clobert, F. Dziock *et al.* (2009) Advantages of Volunteer-Based Biodiversity Monitoring in Europe. Conservation Biology 23(2):307–316.

10 Bison M., N. G. Yoccoz, B. Z. Carlson and A. Delestrade (2019) Comparison of budburst phenology trends and precision among participants in a citizen science program. International Journal of Biometeorology 63:61–72.

11 Meentemeyer R. K., M. A. Dorning, J. B. Vogler, D. Schmidt and M. Garbelotto (2015) Citizen science helps predict risk of emerging infectious disease. Frontiers in Ecology and the Environment 13(4):189–194.

12 Kerbiriou C., C. Azam, J. Touroult, J. Marmet, J.-F. Julien and V. Pellissier (2018) Common bats are more abundant within Natura 2000 areas. Biological Conservation 217:66–74.

13 Schuster R., S. Wilson, A. D. Rodewald, P. Arcese, D. Fink, T. Auer and J. R. Bennett (2019) Optimizing the conservation of migratory species over their full annual cycle. Nature Communications 10:1754.

Tableau 1. Exemples de projets de SRP sur la biodiversité sollicitant les citoyens français pour la collecte de données de masse.

PROJET (DÉBUT)	OBJECTIF	PUBLIC VISÉ	MÉTHODES	TEMPORALITÉ	ÉTENDUE GÉOGRAPHIQUE
SPIPoll (2010)	Suivi photographique des insectes visitant les plantes	Tout public	Observations protocolées, pas de plan d'échantillonnage	Long terme	Nationale (France)
Observatoire des saisons (2006)*	Suivi de la phénologie des plantes et des animaux en lien avec le changement climatique	Tout public	Observations protocolées	Long terme	Nationale (France)
CiTique (2017)*	Développement d'outils de prévention des maladies transmises par les tiques	Tout public	Observations opportunistes, collecte de matériel biologique	Long terme	Nationale (France)
Pl@ntNet (2009)*	Cartographie de l'aire de répartition des plantes	Tout public	Observations opportunistes	Long terme	Mondiale
Vigil'encre (2020)*	Suivi sanitaire des châtaigniers en lien avec la maladie de l'encre	Professionnels et propriétaires forestiers	Observations opportunistes	Long terme	Nationale (France)
Tree Bodyguards (2018)*	Évaluation du contrôle biologique des herbivores du chêne par les prédateurs	Scolaires	Observations protocolées, collecte de matériel	Court terme	Européenne
Sauvages de ma rue (2012)	Suivi de la biodiversité végétale en milieu urbain	Tout public	Observations protocolées, pas de plan d'échantillonnage	Long terme	Nationale (France)
Oiseaux des jardins	Suivi spatio-temporel des espèces d'oiseaux dans les jardins	Tout public	Observations protocolées, site choisi par l'observateur	Long terme	Nationale (France)
Observatoire des sentinelles du climat	Suivi de la phénologie de 18 espèces communes (plantes, animaux, insectes).	Tout public	Observations protocolées	Long terme	Régionale (Nouvelle-Aquitaine)
Faune France	Observations naturalistes	Naturalistes	Observations opportunistes	Long terme	Nationale (France)
Suivi Temporel des Oiseaux Communs **STOC** (1989)	Suivi des populations d'oiseaux	Naturalistes	Observations protocolées, plan d'échantillonnage imposé	Long terme	Nationale (France)
Observatoire Agricole de la Biodiversité (2009)	Suivi de l'impact des pratiques agricoles sur la biodiversité	Agriculteurs	Observations protocolées, site choisi par l'observateur	Long terme	Nationale (France)
Vigie-Nature Ecole (2010)	Suivi de la biodiversité commune, initiation à la démarche scientifique	Scolaires	Observations protocolées, site choisi par l'observateur	Long terme	Nationale (France)
EPOC (2017)	Suivi des oiseaux communs	Naturalistes	Observations protocolées, site choisi par l'observateur	Long terme	Nationale (France)

* projets dans lesquels INRAE est coordinateur ou partenaire.

d'observation des chercheurs dans l'espace et, le temps. Il devient alors possible d'obtenir des informations sur la distribution ou la phénologie des espèces sur l'ensemble de leur aire de répartition, ou de synchroniser une campagne d'observation sur une grande aire géographique[7,10,11]. De plus, la recherche en écologie peut impliquer la répétition de tâches extrêmement chronophages mais ne requérant pas ou que peu d'expertise. La participation du public à ces tâches est un gain de temps substantiel pour les scientifiques, à un coût dérisoire par rapport à ce qu'il serait s'il fallait avoir recours à des collecteurs de données rémunérés[14].

Ce serait sortir du cadre de cet article que de s'attarder sur les bénéfices individuels et sociétaux que procure la participation du public aux programmes de SRP. Il est tout de même utile de les mentionner brièvement, parce que la (re)connaissance de ces bénéfices oriente la manière dont sont conçus les programmes de SRP. Les programmes de SRP fondés sur la collecte de données de masse visent à l'amélioration des connaissances scientifiques et de la capacité à lire, écrire et comprendre la science (i.e. la littératie scientifique) au niveau individuel et au niveau de la communauté. Indissociablement, c'est aussi la transformation de la relation entre le public et la nature au sens large qui est en jeu[15].Couplés aux objectifs scientifiques de la collecte de données par le grand public, ces aspects transformatifs aux niveaux individuel et collectif peuvent également contribuer à orienter les politiques publiques. À titre d'exemple, le Nutri-Score que l'on retrouve sur les emballages de nos produits alimentaires est une mesure de santé publique soutenue par les résultats scientifiques de l'étude NutriNet-Santé basée sur les pratiques de consommation autodéclarées d'une large cohorte d'adultes volontaires[16]. De manière similaire, le Farmland Bird Indicator, basé sur des données de suivis participatifs dans plusieurs pays européens, dont la France avec le STOC, est un indicateur officiel de la communauté européenne pour évaluer l'impact des politiques agricoles sur la biodiversité[17].

Des réticences des deux côtés

La collecte massive de données par le public au travers des SRP apparaît, sur le principe, comme une initiative "gagnant-gagnant". Pourtant, elle se heurte à un certain nombre de réticences ou de critiques synthétisées dans le rapport de Houllier et Merilhou-Goudard[18] sur les sciences participatives en France. Sans chercher à être exhaustif, retenons que des réticences quant à l'utilisation des données collectées en masse par le grand public existent des deux côtés : chez les chercheurs et chez le public. Les enquêtes menées auprès du public pointent du doigt (i) la crainte d'une instrumentalisation des données récoltées à des fins idéologiques ou politiques, (ii) une autocensure (« je ne suis pas assez compétent pour identifier correctement une espèce »), (iii) un renoncement face à l'ambiguïté ou la lourdeur des protocoles (ou au coût de sa mise en œuvre), (iv) un désintérêt pour la "biodiversité ordinaire" ainsi que (v) une déception quant à la nature des résultats obtenus, si les attentes initiales étaient trop fortes[18,19]. Certains scientifiques adoptent une attitude circonspecte, voire hostile, vis-à-vis de la qualité des données générées par les sciences et recherches participatives en général, et par le crowdsourcing en particulier. Plusieurs raisons sont invoquées[18,20,21] : (i) certaines données sensibles ne devraient pas être acquises par le grand public, et encore moins lui être accessibles (par exemple les signalements d'espèces rares ou en danger) ; (ii) il peut y avoir un risque d'instrumentalisation ou de sabotage des données si elles sont préférentiellement acquises par des individus ou des groupes d'individus porteurs d'intérêts politiques ; (iii) tous les scientifiques ne sont pas à l'aise à l'idée d'interagir avec le grand public, et anticipent des difficultés à gérer un pro-

14 Levrel H., B. Fontaine, P.-Y. Henry, F. Jiguet, R. Julliard, C. Kerbiriou and C. Denis (2010) Balancing state and volunteer investment in biodiversity monitoring for the implementation of CBD indicators: A French example Ecological Economics 69(7):1580–1586.

15 Deguines N., K. Princé, A.-C. Prévot and B. Fontaine (2020) Assessing the emergence of pro-biodiversity practices in citizen scientists of a backyard butterfly survey. Science of The Total Environment 716:136842.

16 Julia C., S. Péneau, C. Buscail, R. Gonzalez, M. Touvier, S. Hercberg and E. Kesse-Guyot (2017) Perception of different formats of front-of-pack nutrition labels according to sociodemographic, lifestyle and dietary factors in a French population: cross-sectional study among the NutriNet-Santé cohort participants. BMJ Open. 2017 Jun 15;7 doi: 10.1136/bmjopen-2017-016108. PMID: 28619781; PMCID: PMC5726055.

17 Scholefield P., L. Firbank, S. Butler, K. Norris, L. M. Jones and S. Petit (2011) Modelling the European Farmland Bird Indicator in response to forecast land-use change in Europe. Ecological Indicators 11(1):46–51.

18 Houllier F. and J.-B. Merilhou-Goudard (2016) Les sciences participatives en France : Etat des lieux, bonnes pratiques et recommandations. Paris, Mission Sciences participatives Rapport pour le MENESP, 123 p.

19 Hobbs, SJ and White, PCL. 2012. Motivations and barriers in relation to community participation in biodiversity recording. Journal for Nature Conservation, 20(6): 364–373.

20 Burgess H., L. DeBey, H. Froehlich, N. Schmidt *et al.* (2016) The science of citizen science: Exploring barriers to use as a primary research tool. Biological Conservation 208, 113-120.

21 Law E., K. Z. Gajos, A. Wiggins, M. L. Gray and A. Williams (2017) Crowdsourcing as a Tool for Research: Implications of Uncertainty,1544–1561, Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. Association for Computing Machinery, Portland, Oregon, USA.

jet de recherche aux trop nombreux participants mal identifiés d'un point de vue institutionnel ; (iv) il y a un doute sur la qualité des données acquises par des non-experts.

Plusieurs auteurs ont analysé ces différents points de blocage et émis des recommandations pour les lever[8,22,23]. Nous y renvoyons le lecteur intéressé. Par la suite, nous nous focaliserons sur le dernier point : la qualité présumée faible des données.

Le grand public est-il fiable pour acquérir des données de qualité ?

Qu'est-ce qu'une donnée de qualité ?

La qualité des données acquises peut être vue selon différents critères établis par Pipino *et al.*[24]. En effet, assurer une qualité suffisante pour répondre à une question scientifique représente des coûts, avant et après la collecte. Initier et conduire un projet de SRP impose d'anticiper ces coûts. Dans le cadre des SRP sur la biodiversité, le coût associé à l'acquisition de données de qualité peut se décomposer ainsi :

- le coût de développement d'une infrastructure (site web, application smartphone, etc.) pour la digitalisation,
- les coûts liés à l'accessibilité des données (où sont-elles stockées ?), à leur révision, leur validation et leur protection de manière à assurer leur fiabilité et leur pérennité.

Les coûts liés à la logistique du stockage et de l'accessibilité des données dépendent fortement de leur quantité attendue (et donc de l'étendue spatiale et temporelle du projet, ainsi que des efforts de communication faits pour le publiciser) et du niveau de précision attendu pour répondre aux questions posées (par exemple, des identifications d'organismes au niveau de la famille, du genre, de l'espèce). Ainsi, l'observatoire des bourdons et papillons des jardins a permis de mesurer l'impact de l'utilisation de pesticides sur les insectes floricoles dans les jardins privés en utilisant des identifications grossières (papillons et bourdons), mais suffisantes, pour obtenir des résultats fiables, qu'il aurait été compliqué d'obtenir autrement[6]. Ici, le niveau de précision requis est suffisamment bas pour que les compétences grand public suffisent à procurer des données fiables. En effet, la fiabilité est une composante déterminante de la qualité des données d'un projet de SRP, car elle assure la validité des résultats scientifiques qui en sont issus. Encore faut-il s'entendre sur ce qu'est une donnée fiable.

Qu'est-ce qu'une donnée fiable ?

La fiabilité des données concerne l'incertitude sur la valeur de la donnée fournie par l'observateur. Dans le cadre des SRP sur la biodiversité, il s'agit notamment de l'identification du groupe taxonomique (famille, espèce), mais aussi, parfois, du nombre d'individus, du lieu et de la date d'observation. L'incertitude d'identification et le nombre d'individus observés sont fortement liés au niveau d'expertise de l'observateur. Les incertitudes géographiques et temporelles sont, quant à elles, fortement conditionnées par la précision des outils connectés employés sur le terrain pour la collecte de données ainsi que par l'accès au réseau. Une donnée de biodiversité fiable est donc une donnée pour laquelle l'identification taxonomique est correcte et contextualisée, c'est-à-dire géoréférencée et temporisée avec un niveau de précision suffisant au regard de la question scientifique posée. Cela est conditionné par le protocole de collecte de données ainsi que par le plan d'échantillonnage spatial et temporel des données[25]. Ces deux aspects peuvent, ou non, être définis en amont par les scientifiques en charge du projet.

Il faut donc garder à l'esprit que chaque type de données a sa spécificité et ne peut permettre de répondre qu'à un certain nombre de questions scientifiques, en général, prédéterminées par le protocole ou le contexte d'acquisition (même si la question peut émerger, a posteriori, de l'exploration des données). Ce n'est donc pas tant la qualité de la donnée qui est critique dans le cadre des SRP que la connaissance que les scientifiques ont sur cette qualité[26], et que ce qu'ils en font.

22 Crowston K., E. Mitchell and C. Østerlund (2019) Coordinating Advanced Crowd Work: Extending Citizen Science. Citizen Science: Theory and Practice 4:16.

23 Serret H., N. Deguines, Y. Jang, G. Lois and R. Julliard (2019) Data Quality and Participant Engagement in Citizen Science: Comparing Two Approaches for Monitoring Pollinators in France and South Korea. Citizen Science: Theory and Practice 4(1):22.

24 Pipino L. L., Y. W. Lee and R. Y. Wang (2002) Data quality assessment. Communications of the ACM 45(4):211–218.

25 Miller D. A. W., K. Pacifici, J. S. Sanderlin and B. J. Reich (2019) The recent past and promising future for data integration methods to estimate species' distributions. Methods in Ecology and Evolution 10(1):22–37.

26 Shirk J. L. and R. Bonney (2018) Scientific impacts and innovations of citizen science. Citizen Science: Innovation in Open Science, Society and Policy. UCL Press. UCL Press, 41-51.

Le grand public fournit des données plus ou moins fiables

Plusieurs études ont comparé les données récoltées par les citoyens volontaires à des données de référence ou à des données acquises dans les mêmes conditions par les scientifiques eux-mêmes[27,28]. Elles fournissent des conclusions contrastées quant à la confiance à accorder aux données issues de sciences citoyennes. Aceves-Bueno *et al.*[29] ont récemment réalisé une synthèse de 63 articles ayant directement abordé cette question. Leur étude fait notamment ressortir que : (i) sur la masse des articles scientifiques s'appuyant sur les SRP, peu ont cherché à évaluer la qualité des données, et moins encore ont défini les critères caractérisant une donnée de qualité ; (ii) à peine plus de la moitié des articles évalués fournissent des données de qualité suffisante (selon les critères des auteurs). À noter toutefois que la synthèse d'Aceves-Bueno *et al.*[29] ne se limitait pas aux données issues de crowdsourcing mais prenait aussi en compte un nombre important de "petits" projets n'impliquant qu'une dizaine de participants. Bien que la problématique de la qualité des données soit différente d'un projet à l'autre, le constat peut sembler alarmant en première instance, mais l'examen détaillé des sources de variabilité entre études permet d'établir plusieurs recommandations permettant de réduire les sources d'erreurs dans les données acquises par le grand public[28].

Comment les chercheurs peuvent-ils s'assurer de la qualité des données fournies en masse par le grand public ?

Avant d'acquérir son statut de "donnée", l'observation brute faite par le citoyen passe au travers de plusieurs étapes de préparation, vérification et "réparations" avant de pouvoir être considérée comme fiable et informative, donc exploitable à des fins de production de savoirs et de connaissances scientifiques (voir l'infographie, Figure 1).

Le protocole standardise les données

Données standardisées - Les données issues des programmes de SRP sont dites standardisées quand elles sont collectées en respectant un protocole assurant qu'elles soient comparables dans le temps et dans l'espace. Ce protocole implique le respect de contraintes sur la pression d'observation, les techniques et le matériel utilisé, et la date, le lieu et les conditions d'observation. Pour le matériel, par exemple, il suffit d'un appareil photo (ou d'un smartphone) pour participer au SPIPOLL (Tableau 1). Au contraire, le projet *Tree Bodyguards*, qui invite les élèves des écoles primaires et secondaires à fabriquer et installer

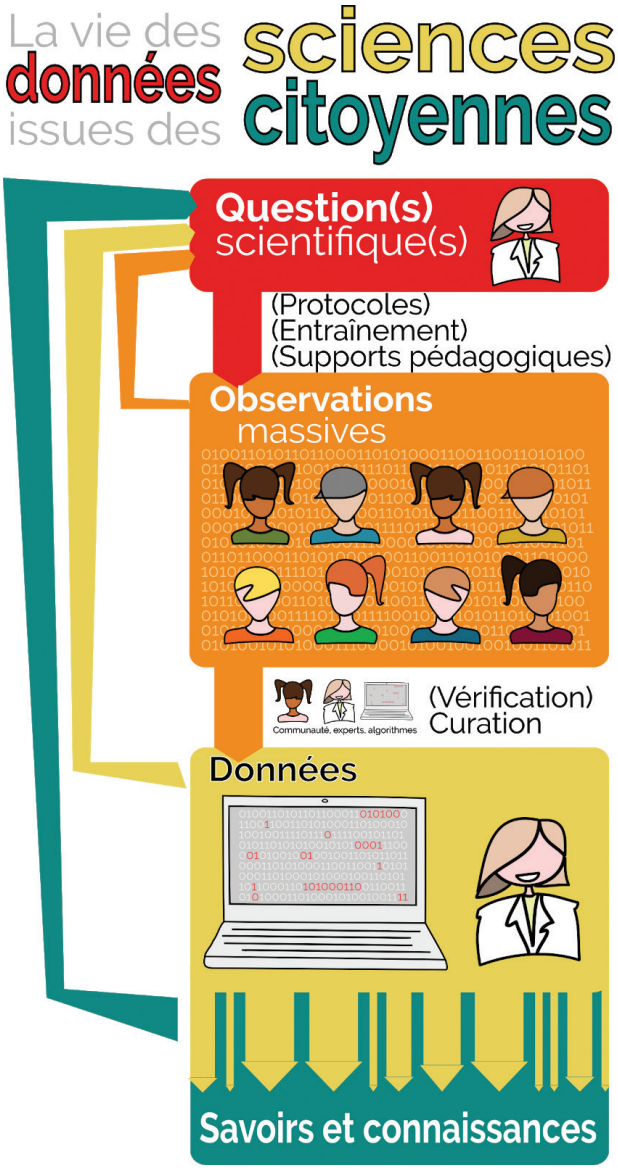

Figure 1. Le cycle de vie des données de *crowdsourcing* acquises dans le cadre des sciences et recherches participatives. Toutes les étapes présentées dans cette infographie ne se retrouvent pas dans tous les programmes de sciences et recherches participatives basées sur le *crowdsourcing* (vérification, protocole, entraînement, supports pédagogiques peuvent manquer).

27 Castagneyrol B., E. Valdés-Correcher, A. Bourdin, L. Barbaro, O. Bouriaud, M. Branco *et al.* (2020) Can School Children Support Ecological Research? Lessons from the Oak Bodyguard Citizen Science Project. Citizen Science: Theory and Practice 5(1),1-11.

28 Balázs, Mooney, Nováková, Bastin, Arsanjani, 2021. Data quality in Citizen Science. In: Vohland K. *et al.* (eds) The Science of Citizen Science. Springer, Cham. https://doi.org/10.1007/978-3-030-58278-4_8.

29 AcevesBueno E, A. S. Adeleye, M. Feraud, Y. Huang, M. Tao, Y. Yang, et S. E. Anderson (2017) The Accuracy of Citizen Science Data: A Quantitative Review. The Bulletin of the Ecological Society of America 98(4):278–290.

des chenilles en pâte à modeler dans les arbres pour étudier la prédation par les oiseaux, fournit un matériel spécifique. Pour le STOC, il est interdit d'utiliser des jumelles pour repérer les oiseaux.
Le plan d'échantillonnage peut imposer les dates et sites d'acquisition des données ou, au contraire, avoir une approche opportuniste. Par exemple, le STOC repose sur un plan d'échantillonnage spatial et temporel très strict, avec des sites tirés aléatoirement (afin d'assurer une bonne représentativité des habitats). Au contraire, le SPIPOLL ou l'EPOC (Tableau 1) n'ont pas de plan d'échantillonnage, les participants peuvent effectuer des sessions où bon leur semble. On parle alors de données semi-structurées (présence d'un protocole, mais sans plan d'échantillonnage). Cas intermédiaire, certains programmes laissent aux observateurs le libre choix de leur site, mais ils doivent ensuite revenir régulièrement sur ce site (par exemple, Oiseaux des Jardins, dans lequel les observateurs collectent des données sur les oiseaux visitant leur jardin).

Le protocole peut également laisser les citoyens libres de participer quand bon leur semble, ou imposer des observations à date fixe, selon des conditions particulières. Par exemple, pour le STERF, les papillons sont observés entre 10h et 18h, avec une couverture nuageuse d'au maximum 75 % et sans pluie, un vent inférieur à 30 km/h et une température d'au moins 13°C, si le temps est ensoleillé ou faiblement nuageux, ou d'au moins 17°C si le temps est nuageux.

Données opportunistes - À l'échelle mondiale, la majorité des projets de SRP repose sur l'acquisition de données opportunistes, grandement facilitée par le développement des outils numériques nomades. Les données opportunistes sont, par définition, collectées au bon vouloir des observateurs. Les scientifiques pilotant le projet n'ont pas de connaissance a priori ni aucun moyen de contrôle sur l'origine et le volume de données acquises de manière opportuniste. Par exemple, les utilisateurs de l'application Pl@ntNet (Tableau 1) ne font remonter des observations d'occurrence que dans les secteurs qu'ils décident de visiter, de sorte que des secteurs géographiques entiers sont sous-échantillonnés, alors que d'autres sont sur-représentés. De telles données sont donc lacunaires et biaisées, pour certaines questions de recherche, et cela peut limiter leur exploitation. En effet, de très nombreuses données de présences seules opportunistes ne nous informent pas sur l'absence avérée d'une espèce là où aucune observation n'a été rapportée. Cela peut arriver si une zone n'a tout simplement pas été prospectée, ou si les observateurs ont rapporté leurs observations de manière sélective (par exemple, en ne signalant que la présence d'espèces rares ou remarquables, délaissant les espèces communes).

Qu'elles soient opportunistes ou protocolées, les données sont en général générées pour répondre à une question précise. C'est la question qui détermine le niveau de précision dans les observations demandées aux citoyens et les contraintes imposées par le protocole. La réalisation d'un inventaire exhaustif de la diversité des hyménoptères d'une région donnée requiert des compétences hors de portée des amateurs, mais pour mesurer l'impact de l'urbanisation sur les communautés d'insectes floricoles[5,30] ou celui des pesticides dans les jardins privés pour les papillons et bourdons, il devient acceptable de ne pas chercher à avoir une résolution taxonomique au niveau de l'espèce.

Au final, la présence d'un protocole contraignant facilite l'analyse des données, mais peut limiter le nombre de volontaires susceptibles de les acquérir de manière fiable. Au contraire, l'absence de protocole impose le recours à des outils d'analyse complexes permettant de tenir compte des biais d'observation. Ce format peu exigeant de collecte de données ouvre néanmoins la porte à un maximum de contributeurs, et peut ainsi permettre de couvrir des espaces et fenêtres temporelles plus étendus. Les données opportunistes sont également plus à même de répondre, a posteriori, à des questions scientifiques qui n'auraient pas été anticipées (i.e., une forme de sérendipité). Par exemple, la mise en place, par la LPO, des bases Faune France avait pour principal objectif d'accumuler des données permettant d'étudier la répartition spatio-temporelle des oiseaux, sans question bien identifiée au départ.

Les données sont d'autant plus précises que les volontaires gagnent en expérience

Les SRP ont un objectif transformatif d'acquisition de connaissances et de développement de littératie scientifique auprès du grand public[8,31]. Le gain de connaissance est une des sources de motivation que l'on retrouve chez les citoyens volontaires. La prise en compte de ce facteur dans

30 Deguines N., M. de Flores, G. Loïs, R. Julliard and C. Fontaine (2018) Fostering close encounters of the entomological kind. Frontiers in Ecology and the Environment, Ecological Society of America, 202–203.

31 Aristeidou M. and C. Herodotou (2020) Online Citizen Science: A Systematic Review of Effects on Learning and Scientific Literacy. Citizen Science: Theory and Practice 5(1),1-12.

la préparation des projets de SRP permet non seulement de renforcer la participation du public, mais augmente aussi significativement le degré d'expertise des participants et, par là même, la qualité des données. Par exemple, Ratniek *et al.*[32] ont comparé la capacité de trois groupes de volontaires à identifier les insectes visitant le lierre selon qu'ils avaient reçu un entraînement à base de fiches descriptives et de présentations powerpoint ou un entraînement sur le terrain avec des scientifiques. Ils ont montré que la précision dans les identifications d'insectes était positivement associée à l'intensité de l'entraînement reçu. De la même manière, les compétences taxinomiques des observateurs du SPIPOLL augmentent avec le nombre d'observations[30].

Pour permettre au grand public d'améliorer sa précision d'identification, certains projets proposent l'assistance par des algorithmes d'identification automatisés. C'est le parti pris par le projet Pl@ntNet qui intègre la reconnaissance automatisée d'images de plantes à son application mobile depuis ses débuts, en 2013[33]. Cette approche est motivée par le fait qu'il existe plus de 300 000 espèces de plantes connues à travers le monde, de nombreuses confusions étant possibles, même à une échelle locale, alors que l'expertise botanique se raréfie. L'algorithme de Pl@ntNet permet d'aiguiller l'observateur jusqu'au bon genre, voire la bonne espèce, dans la vaste majorité des cas pour les zones les plus prospectées. Cet algorithme s'améliore constamment[34] et, notamment, grâce l'enrichissement de sa base d'apprentissage par les identifications validées de la communauté. Ceci contribue à disposer d'algorithmes de plus en plus performants, dont les capacités sont de plus en plus proches des personnes les mieux formées dans ce domaine. De tels systèmes automatisés sont déjà disponibles et fonctionnels gratuitement pour d'autres types d'organismes, comme les oiseaux[35], et peuvent s'intégrer à des SRP plus spécifiques ou de moins grande ampleur, avec un double objectif : faciliter les apprentissages chez les volontaires et (ainsi) renforcer la qualité des données.

Les observations peuvent être vérifiées avant d'être utilisées

Plusieurs procédures permettent de vérifier la pertinence des observations brutes. Selon les projets, cette opération peut être réalisée par les scientifiques ou des experts, par les citoyens observateurs eux-mêmes ou être automatisée. Cette étape de validation a pour objectif de supprimer les observations aberrantes avant que les données ne soient soumises à l'analyse.

Par exemple, dans le cadre du projet *Tree Bodyguards*, les données sont directement acquises par les scientifiques en charge du projet, à partir du matériel fourni par les participants[27,36]. À son lancement, les observations du SPIPOLL étaient d'abord filtrées par les observateurs eux-mêmes qui ne transmettaient que les photographies d'insectes répondant à un cahier des charges précis[23,37]. Les données étaient ensuite validées par des experts (naturalistes de l'Office Pour les Insectes et leur Environnement, OPIE). Un système analogue de validation par des experts (botanistes de la communauté Tela Botanica) a permis le lancement de l'application Pl@ntNet[38].

Ce système de validation a permis un gain en compétence des utilisateurs (SPIPOLL) et la constitution de bases de données photographiques alimentant des algorithmes d'intelligence artificielle (Pl@nNet), de sorte que la validation initiale des données par des experts a été peu à peu complétée par un autre système[33]. Depuis 2019, la validation des observations du SPIPOLL est assurée par les participants eux-mêmes : pour être considérée comme valide, une identification doit avoir obtenu un score de 3 points. Un point est obtenu lorsqu'un participant (autre que celui qui a fait la photo) valide l'identification. Chaque participant ne peut valider qu'une fois une photo donnée. Un participant peut également enlever un point à l'identification, s'il considère que celle-ci est fausse (même si elle avait déjà

32 Ratnieks F. L. W., F. Schrell, R. C. Sheppard, E. Brown, O. E. Bristow and M. Garbuzov (2016) Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods in Ecology and Evolution 7(10):1226–1235.

33 Joly A., P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, A. Affouard, J. Carré, J.-F. Molino, N. Boujemaa and D. Barthélémy (2016) A look inside the Pl@ntNet experience. Multimedia Systems 22:751–766.

34 Goëau H., P. Bonnet and A. Joly (2017) Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). Conference and Labs of the Evaluation Forum. Sep 2017 Dublin, Ireland.

35 Van Horn G., S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona and S. Belongie (2015) Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. 595–604,Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway, NJ.

36 Castagneyrol B., E. Valdés-Correcher, M. Kaennel Dobbertin and M. Gossner (2019) Predation assessment on fake caterpillars and leaf sampling: Protocol for partner schools. https://doi.org/10.17504/protocols.io.42pgydn.

37 Deguines N., R. Julliard, M. de Flores and C. Fontaine (2012) The Whereabouts of Flower Visitors: Contrasting Land-Use Preferences Revealed by a Country-Wide Survey Based on Citizen Science. https://doi.org/10.1371/journal.pone.0045822.

38 Barthélémy D., N. Boujemaa, J.-F. Molino, A. Joly, H. Goëau, V. Bakić, S. Selmi, J. Champ, J. Carre, M. Chouet, A. Perronnet, C. Vignau, S. Dufour-Kowalski, A. Affouard, J. Barbe and P. Bonnet (2014) Pl@ntNet, une plate-forme innovante d'agrégation et partage d'observations

été validée). Ce système permet de s'affranchir des experts, en assurant une très bonne qualité des identifications, et en favorisant des interactions entre les participants, bénéfiques au dynamisme du programme. Dans le cadre de Pl@ntNet, toute observation est accompagnée d'une ou plusieurs images de la plante et d'une localisation dans la plupart des cas. Ces observations brutes réalisées au travers de l'application passent ensuite par un système de révision collaborative faisant intervenir toute la communauté des utilisateurs[38]. Le système de révision se décline ainsi en une révision de la détermination taxonomique proposée, notamment via le site IdentiPlante, et en une estimation de la qualité visuelle des images illustrant l'observation, initialement effectuée via le site PictoFlora. L'observateur a alors la possibilité d'indiquer son accord, désaccord ou de proposer une identification alternative. Chaque groupe d'images reçoit donc un certain nombre de votes, pondérés par le niveau d'expertise des votants, lequel s'accroît avec le nombre et la diversité de leurs observations précédemment validées. Ainsi, la validation des observations est permise par la communauté d'expertise botanique citoyenne. Toutes les données validées sont ensuite partagées avec la communauté internationale à travers différents réseaux scientifiques, le plus important d'entre eux étant le réseau du usagers de la plateforme du GBIF.

Toutes les données générées par les programmes de SRP ne sont pas nécessairement vérifiées ou vérifiables. Les données de la majorité des programmes de Vigie-Nature sont déclaratives et donc invérifiables. Le choix a donc été fait de ne pas passer par des étapes de validation des données une à une. Ces programmes génèrent, en effet, des centaines de milliers de données (voire davantage) qu'il serait trop coûteux de vérifier. De plus, ces programmes s'intéressant aux espèces communes, la majorité des données sont invérifiables : comment savoir si une fauvette à tête noire, identifiée par un participant au STOC, n'a pas été confondue avec une fauvette des jardins ? En effet, les deux espèces sont communes, ont une large aire de répartition et leur présence sur un site est donc vraisemblable autant pour l'une que pour l'autre. Seuls les programmes dans lesquels les données ne sont pas uniquement déclaratives, mais reposent sur des supports numériques, font l'objet de validation : validation par les pairs, pour les photos du SPIPOLL, identification automatique et validation par les observateurs et les experts, pour les sonogrammes issus de Vigie-Chiro.

Un traitement statistique approprié prend en compte les spécificités de chaque type de données

Pour répondre à une question scientifique, la modélisation statistique simplifie et formalise mathématiquement le lien entre les variables mesurées (les données) par des équations qui les relient. On fait alors intervenir des facteurs aléatoires ou déterministes représentant les phénomènes d'observation qui peuvent perturber ou biaiser les valeurs mesurées par rapport à la réalité. Prenons un exemple. Le projet *Tree bodyguards* cherche à relier la température et le service de régulation des insectes herbivores par les oiseaux. Pour cela, un réseau européen de scientifiques et d'écoliers a déployé de fausses chenilles en pâte à modeler dans des chênes (20 chenilles par arbre), et a dénombré le nombre de fausses chenilles présentant des coups de bec. Pour chaque chêne, deux données étaient disponibles : le pourcentage de chenilles attaquées (P), et la température annuelle moyenne du site (T). En faisant l'hypothèse que la prédation par les oiseaux est déterminée par la température, alors P et T peuvent être reliés par l'équation :

$$P = a \times T + b + \varepsilon$$

P et T sont les données (connues), a et b les paramètres de l'équation, et ε l'erreur, soit toute la variation de P qui n'est pas expliquée par T et qui inclut l'erreur de mesure ou d'observation. Tout l'enjeu de la modélisation statistique consiste à attribuer une valeur aux paramètres a et b, à quantifier l'incertitude autour de ces paramètres, et à quantifier ε. Un enjeu majeur de l'analyse des données issues des SRP basées sur le crowdsourcing – comme d'ailleurs des données acquises par des experts – consiste à contrôler et modéliser ε. Or, plusieurs facteurs peuvent influencer l'incertitude autour des relations entre les variables. Nous développerons deux exemples : les erreurs de mesures et le caractère opportuniste des données.

Erreurs sur la mesure - Nous avons vu, plus haut, qu'elles peuvent être en partie contrôlées, mais qu'elles sont inévitables. Toutefois, elles peuvent parfois être ignorées, d'autant plus que les progrès technologiques (par exemple la précision des GPS) rendent possible la collecte de données de plus en plus précises à des coûts de moins en moins élevés, de sorte que certaines données collectées aujourd'hui par un citoyen moyen peuvent être plus précises que les mêmes données collectées il y a 10 ans par un expert.

botaniques,191–197, In: N. R. Rakotoarisoa, S. Blackmore and B. Riera, editors. International Conference 'Botanists of the Twenty-first Century' Sep 2014, UNESCO, Paris, France.

L'objectif de certains programmes, comme l'observatoire des papillons des jardins (OPJ), est le suivi à long terme d'un groupe d'organismes : les observations sont recensées de la même manière par tous les participants, aux même dates, selon un protocole précis qui ne change pas. On peut faire l'hypothèse que le taux d'erreur, qui n'est jamais nul, ne change pas avec le temps. On décide de l'accepter en rendant la vérification des identifications une à une inutile. Les données issues de l'OPJ, ont par exemple, permis de montrer que l'urbanisation réduit la diversité et l'abondance des papillons, mais que ces effets peuvent être contrebalancés à l'échelle locale par la plantation et l'entretien de plantes nectarifères[39]. Dans un tel cas, l'impact de l'erreur ou des biais d'identification sur les résultats peut être supposé négligeable.

Caractère opportuniste des observations - Si l'on ne peut pas faire l'hypothèse a priori que les erreurs d'observations ne sont pas constantes dans l'espace et le temps, alors ε doit faire l'objet d'un effort de modélisation supplémentaire. C'est le cas lorsqu'il s'agit de comparer des abondances ou des diversités d'espèces entre sites (habitats, régions, pays, biomes) et que la fiabilité des données est imparfaite, inégale et surtout inconnue.

Il existe des protocoles de collecte qui permettent d'estimer les probabilités de détection des divers observateurs et les probabilités d'occupation par site en parallèle, en vue de produire des cartes de répartition ou d'estimer des préférences environnementales. C'est le cas des modèles d'occupation en détection imparfaite. Ils se basent sur des visites répétées d'un site, sur une même période de temps, par plusieurs observateurs, avec connaissance de l'effort d'observation. En utilisant d'importantes quantités de ces données, la modélisation de la variabilité des capacités de détections en fonction des observateurs, des conditions et de l'effort d'observation, permet de corriger ces facteurs de biais, et de retrouver la répartition d'une espèce de manière fiable. Ce type de modélisation a été récemment employé pour étudier la recolonisation du loup gris, en France, à partir d'observations citoyennes[40].

Pour les données opportunistes, dites de présence-seule, c'est-à-dire d'observations ponctuelles et localisées de présence d'une espèce, l'information sur le processus d'observation est quasi-nulle : on ne connaît ni le temps ni l'effort d'échantillonnage consacrés, pas plus que l'intérêt relatif porté par l'observateur aux espèces rares et communes. Or, tous ces facteurs peuvent varier énormément à l'échelle d'un projet de SRP[41]. De plus, pour les données en présence seule, l'absence de donnée ne peut pas être considérée comme une donnée d'absence : il peut ne pas y avoir d'information sur la présence du tournepierre à collier au sommet du Néouvielle parce qu'il ne s'y trouve pas du fait de son écologie, ou parce qu'aucun observateur n'a prospecté ce site. Enfin, par essence, ce type de données ne permet pas d'obtenir l'information d'abondance absolue[42], et comparer la concentration des occurrences d'une zone à une autre n'est pas non plus une bonne mesure d'abondance relative.

Les modélisateurs ont développé plusieurs stratégies pour contourner ces problèmes, faisant souvent appel à des hypothèses sur le processus d'observation. Il est, par exemple, possible de modéliser l'effort d'échantillonnage comme une fonction de facteurs géographiques ou environnementaux connus (distance aux routes, aux villes, etc.) pour l'estimer conjointement avec la distribution d'une espèce et, ainsi, corriger les biais d'échantillonnages[43]. Par ailleurs, dans certains cas, on peut utiliser une approximation de la distribution de l'effort pour corriger le biais (e.g. un ensemble d'occurrences, toutes espèces confondues[44]). Enfin, une stratégie alternative est de combiner des données plus protocolées, comme des comptages[45], des présences/

39 Fontaine B., B. Bergerot, I. Le Viol and R. Julliard (2016) Impact of urbanization and gardening practices on common butterfly communities in France. Ecology and Evolution 6(22):8174–8180.

40 Louvrier J., C. Duchamp, V. Lauret, E. Marboutin, S. Cubaynes, R. Choquet, C. Miquel and O. Gimenez (2018) Mapping and explaining wolf recolonization in France using dynamic occupancy models and opportunistic data. Ecography 41(4):647–660.

41 Botella C. (2019) October. Statistical methods for spatial plant species distribution modeling based on large masses of uncertain observations from citizen-science programs. Machine Learning [stat.ML] Thèses, Université de Montpellier.

42 Hastie T. and W. Fithian (2013) Inference from presence-only data; the ongoing controversy. Ecography 36(8):864–867.

43 Warton D. I., I. W. Renner et D. Ramp (2013) Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. PLOS ONE https://doi.org/10.1371/journal.pone.0079168.

44 Phillips S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick and S. Ferrier (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19(1):181–197.

45 Giraud C., C. Calenge, C. Coron and R. Julliard (2016) Capitalizing on opportunistic data for monitoring relative abundances of species. Biometrics 72(2) 649–658.

absences ou des données d'occupation sous détection imparfaite, aux données opportunistes[46,47]. Les données protocolées agissent alors comme témoin du biais et permettent de l'éliminer, alors que la masse de données opportunistes augmente la précision de l'estimation. La sélection de différents jeux de données, produits à travers différentes méthodologies sur un même territoire, étant désormais possible à travers le GBIF, la complémentarité de telles données devrait sans nul doute permettre, à court terme, de nouvelles avancées scientifiques.

On voit donc que l'utilisation appropriée du savoir sur le processus d'échantillonnage et la combinaison de données protocolées et opportunistes dans les modèles statistiques permet, dans certains cas, de contrebalancer l'incertitude et l'hétérogénéité des données pour répondre à des questions écologiques ambitieuses.

Conclusion

Les programmes de sciences et recherches participatives (SRP), notamment ceux basés sur l'acquisition de données en masse par le grand public (crowdsourcing), sont devenus un outil incontournable de la recherche sur la biodiversité, tant ils fournissent des données avec une résolution et une profondeur spatiale et temporelle inaccessibles aux chercheurs seuls. Toutefois, parce qu'elles sont acquises par des observateurs dont la motivation est inconnue et l'expertise variable, la qualité des données issues de ces programmes peut être remise en question, ou du moins questionnée a priori. Dans cet article, nous montrons qu'il existe une grande diversité d'approches dans les programmes de SRP, y compris parmi la gamme plus restreinte de programmes fondés sur le crowdsourcing. À partir de quelques exemples représentatifs, nous montrons que les interrogations quant à la qualité des données générées par ces programmes sont légitimes. Il ne s'agit en aucun cas de les nier. Toutefois, en retraçant le cycle de vie des données acquises par les citoyens volontaires, nous montrons que plusieurs niveaux de contrôle de la qualité de la donnée peuvent être appliqués en amont et en aval de leur utilisation pour répondre à une question scientifique. Nous insistons sur le fait que pour peu que la qualité des données soit connue et que les procédures d'analyses et d'inférences la reconnaissent, elle ne doit pas être un frein à l'implémentation des programmes de SRP.

Tout au long de l'article, nous avons vu que l'effort de construction d'un projet de SRP en crowdsourcing répond souvent à un besoin de collecte de données dont la couverture spatiale ou temporelle est inaccessible à un échantillonnage par des experts, et ne nécessite pas une expertise maximale. Aussi, les SRP et le crowdsourcing n'ont pas vocation à remplacer les approches traditionnelles pour l'étude de la biodiversité, pas plus que le recours aux citoyens volontaires pour l'acquisition de données ne doit se substituer à un investissement des universités et instituts de recherche dans la formation et le recrutement de professionnels qualifiés. Au contraire, les SRP et le crowdsourcing doivent être pris pour ce qu'ils sont : un outil complémentaire aux autres approches de la recherche, permettant de répondre efficacement à des questions spécifiques, tout en participant à la sensibilisation et l'éducation du public à la science et à l'environnement. ■

46 Fithian W., J. Elith, T. Hastie and D. A. Keith (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6(4):424–438.

47 Koshkina V., Y. Wang, A. Gordon, R. M. Dorazio, M. White and L. Stone (2017) Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. Methods in Ecology and Evolution 8(4):420–430.