Utilisation des grands jeux de données naturalistes pour répondre aux enjeux scientifiques et sociétaux actuels

Anne-Christine Monnet, Thomas Haevermans, Anne-Sophie Archambeau, Philippe Grandcolas, Roseli Pellens

Chapitre 18

Utilisation des grands jeux de données naturalistes pour répondre aux enjeux scientifiques et sociétaux actuels

Anne-Christine MONNET[1], Thomas HAEVERMANS[1], Anne-Sophie ARCHAMBEAU[2], Philippe GRANDCOLAS[1] et Roseli PELLENS[1]

[1] *Institut de Systématique, Evolution, et Biodiversité, Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, UA, Paris, France*

[2] *IRD, UMS Patrimoine Naturel (OFB-CNRS-MNHN), GBIF France, Paris, France*

18.1. Introduction

Les collections d'histoire naturelle représentent des ensembles d'objets extraordinairement riches et divers. Ces ensembles correspondent également à des jeux de données colossaux qui permettent de répondre à des problématiques scientifiques et sociétales actuelles extrêmement variées. L'informatisation des collections d'histoire naturelle pendant ces dernières décennies, et surtout leur agrégation dans des portails globaux, a donné un accès direct et efficace à la richesse et la diversité de plus de 200 ans de constitution des collections dans plus de 5 000 institutions (muséums d'histoire naturelle, herbiers, centres universitaires) à travers le globe. Un changement d'échelle s'est opéré dans la quantité et la nature des données à disposition des chercheurs conduisant à des changements profonds dans de nombreux domaines scientifiques, notamment la biogéographie, la macroécologie et la biologie de la conservation. Plusieurs questions qui étaient traitées de manière

partielle ou localisée ont pu être revisitées et ré-analysées. De nombreuses questions sociétales contemporaines peuvent ainsi être traitées à des échelles spatiale, temporelle ou taxonomique d'une largeur sans précédent.

Le défi consiste maintenant à s'approprier ces données pour répondre à de nouvelles questions. Or, les données de collections ont la particularité d'avoir été générées avant les questions que se posent leurs utilisateurs d'aujourd'hui. Ces données ont été obtenues à des fins d'inventaire, par opportunité de récolte ou pour répondre à une grande diversité de questions. Elles sont donc potentiellement très hétérogènes en termes de protocoles de récoltes et de nature. Dans ce contexte, comment tirer le meilleur de ces données dans le cadre de nouvelles utilisations actuelles, notamment lorsqu'elles sont plus englobantes en termes d'échelles ? Comment assurer un dialogue mutuellement enrichissant entre programmes de recherches, questions scientifiques, gestion des collections et gestion des bases de données ? Dans ce chapitre nous présentons plusieurs facettes des réutilisations de données issues de collections naturalistes. Nous apportons quelques suggestions pour optimiser les utilisations de données passées et nous assurer que les données issues des études contemporaines puissent être réutilisées dans le futur.

18.2. Mettre à disposition les données : une révolution

Nous sommes entrés dans une ère de numérisation et d'informatisation des collections d'histoire naturelle, précurseuse de la dynamique actuelle de la science ouverte. Les données qui en sont issues sont donc disponibles directement pour un nombre croissant d'utilisateurs. Le but ultime est de rendre disponibles en libre accès les informations associées à tous les spécimens existants dans les collections d'histoire naturelle, qu'il s'agisse des informations de récoltes ou d'utilisations ultérieures. Dans la pratique, la mobilisation des données se fait selon la disponibilité de spécialistes et/ou l'intérêt de chaque institution pour un territoire, une ressource naturelle, ou un groupe d'organisme et, dans des cas plus particuliers, pour répondre à des demandes sociétales spécifiques.

Les données sont souvent mises à disposition *via* des portails d'accès aux données comme le GBIF (*Global Biodiversity Information Facility*) (figure 18.1), par des consortiums de spécialistes des différents groupes taxonomiques (par exemple VertNet[1] pour les vertébrés), ou par la publication de *data papers* des articles permettant de décrire et publier des données d'une étude ou d'une collection naturaliste (Monnet *et al.* 2021). L'utilisateur peut alors exporter directement les données depuis les plates-formes de consultation et de téléchargement mises à

1 www.vertnet.org.

disposition par les fournisseurs de données. Selon les cas, la nature des données sera sensiblement différente, de la donnée brute d'origine jusqu'à la donnée partiellement réinterprétée par les concepteurs de la base. Il faut donc mettre en garde les utilisateurs sur la nature des données mises à disposition, dont la méthode de récolte et de traitement éventuel doit être renseignée.

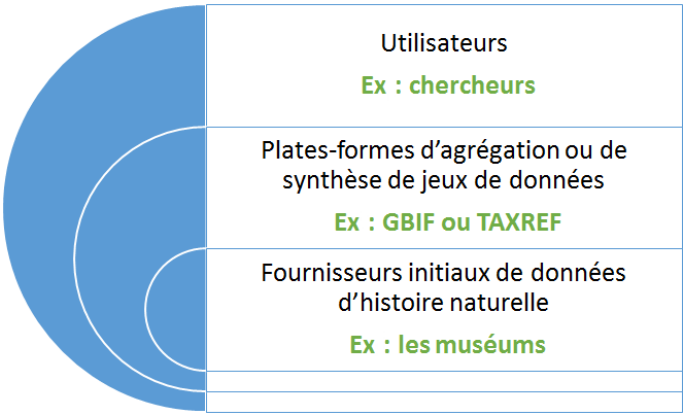

Figure 18.1. *Les différents acteurs, des fournisseurs aux utilisateurs de données en passant par les plates-formes d'agrégation ou de synthèse de données*

Certaines bases comme la *World Flora Online*[2] peuvent proposer des données résultant d'un processus de validation de l'ensemble de la base par un réseau d'expert. D'autres comme *Catalogue of Life* proposent une classification consensuelle issue de différentes sources : si certaines parties des données sont couvertes par des sources différentes, les données jugées les plus pertinentes sont choisies et présentées, au lieu de fournir un consensus des données. Il est donc important de savoir si les données que l'on utilise résultent d'un processus impliquant un examen individuel par un expert, ou bien un processus automatisé de sélection, correction et consensualisation.

La mise en commun et le libre accès aux données permettent non seulement de traiter des questions qui n'auraient pas pu être adressées auparavant, mais aussi de développer une nouvelle dynamique. Des nouvelles méthodes de travail basées sur un partage d'expertise et de connaissances se développent pour traiter ces nouvelles données et leurs avancées, dans le cadre de la science ouverte (voir remarque).

2 www.worldfloraonline.org.

REMARQUE. Estimer le niveau de menace de 300 000 espèces de plantes

Une étude menée par notre équipe s'est donnée pour objectif d'estimer le niveau de menace de l'ensemble des plantes vasculaires (Haevermans *et al.* en préparation). Pour y parvenir, les auteurs ont mobilisé les connaissances fournies par l'Herbier de Paris, mais pas uniquement. Afin d'obtenir toutes les informations nécessaires pour répondre à leur question, plusieurs sources de données ont été réunies :

– les données de l'Herbier du Muséum qui fournissent des informations sur la collecte de plus de six millions de spécimens ;

– les données de la Liste rouge de l'IUCN[3] qui fournissent une évaluation du risque d'extinction des espèces à partir de multiples critères ;

– les données de la liste mondiale de vérification de certaines familles de plantes (*World Checklist of Selected Plant Families*, WCSP ; en anglais[4]) fournie par les Jardins botaniques royaux (*Royal Botanic Gardens*) de Kew qui fournissent des informations géographiques sur les espèces (leurs présences à l'échelle nationale ou infranationale), mais aussi sur leurs traits fonctionnels et leur histoire de vie ;

– la liste d'espèces *The Plant List 2.0*[5] qui constitue la liste d'espèces de plantes la plus complète.

L'étape suivante consiste alors à évaluer la qualité et la pertinence des données pour la question posée. Sur quel référentiel taxonomique s'appuyer ? Est-ce que les noms scientifiques d'espèces sont tous complets et correctement orthographiés ? Combien de données sont manquantes pour chaque information nécessaire ? Est-ce que les codes indiqués pour les zones géographiques correspondent tous au standard utilisé (par exemple, le code ISO des pays ou le code TDWG (Brummitt 2001)) ? Sur des bases de données de si grande taille, ces vérifications ne peuvent se faire visuellement.

Une fois ces premières vérifications effectuées, chacune des bases de données utilisées a été mise en correspondance avec la checklist fournie par Kew qui a servi de référentiel taxonomique pour cette étude.

3 www.iucnredlist.org.

4 www.wcsp.science.kew.org.

5 www.theplantlist.org.

C'est seulement après que le travail d'analyse à proprement parler a pu commencer. Même si ces étapes sont préalablement souvent bien balisées par les gestionnaires de données, ce travail de préparation de données est souvent fastidieux et requiert des compétences techniques avancées. Le temps imparti ne doit ainsi pas être négligé lors de la conception du projet.

Enfin, il est important pour les futurs lecteurs ou utilisateurs de l'étude de pouvoir associer les résultats de l'étude aux versions des bases de données utilisées pour les analyses. Cette bonne pratique assure la reproductibilité des analyses pour des potentielles nouvelles analyses ou ré-analyses. Or, à l'exception de *The Plant List* qui est publiée par versions successives, les bases de données utilisées dans cette étude sont constamment mises à jour. Parce que pour ces bases-ci un identifiant DOI n'est pas fourni lors de l'étape initiale d'extraction, l'idéal consiste alors à associer à la publication de l'étude un répertoire de stockage avec les données utilisées pour archivage.

18.3. Les défis pour les fournisseurs de données

18.3.1. *La lecture des étiquettes ou répertoires*

Informatisation et numérisation demandent principalement à rendre disponibles le spécimen de musée et les informations qui lui sont associées avec des étiquettes ou des répertoires. Les informations concernent traditionnellement les conditions de récoltes des spécimens, mais aussi parfois leurs utilisations ultérieures. Le fournisseur de données doit veiller à la qualité et à la structuration des données et à une documentation convenable des métadonnées. Les informations doivent être les plus proches possibles de la réalité, être comprises par différents utilisateurs, et être utilisables de manière durable par des algorithmes informatiques des portails d'accès.

Le premier défi pour les fournisseurs de données est d'assurer la bonne lecture et extraction des informations des étiquettes et répertoires. Cela peut s'avérer facile avec des documents imprimés, mais ce n'est pas toujours le cas avec des documents plus anciens manuscrits, comportant des toponymies ou des noms rares. Cette lecture peut entraîner des erreurs, et une lecture indépendante par au moins deux personnes permet une prise de décision plus fiable. On ne connaît pas le taux d'erreur ou les problèmes associés à la lecture des étiquettes par une seule personne, la forme la plus courante de saisie de données dans nos institutions. Ceci indique d'une part qu'il est nécessaire de mieux traiter ce problème, et d'autre part que les institutions doivent investir dans des procédures qui assurent une meilleure saisie

originelle des données. Par exemple, afin de contribuer à cette tâche, l'Herbier de Paris a mis en place « Les Herbonautes »[6]. Ce programme destiné à accélérer la lecture des étiquettes grâce à la participation citoyenne propose une méthode répétable de lecture. Chaque étiquette est lue indépendamment par au moins trois personnes, et ses informations sont croisées afin de vérifier le taux de consensus. S'il n'y a pas de consensus, l'étiquette est remise en lecture. Si les problèmes persistent, le cas est pris en main par le responsable de la mission qui cherche la cause du problème et décide de la suite. Le manque de consensus est en soi un indicateur des limites/problèmes dans l'information et il s'est souvent avéré que personne ne pourrait extraire davantage de ces étiquettes (M. Pignal, communication personnelle). L'information en question est donc considérée comme absente.

Les fautes de frappe, l'inversion de caractères, ainsi que la diversité des caractères spéciaux des langues étrangères peuvent donner lieu à des problèmes qui demandent un grand travail de correction par les utilisateurs. Ces problèmes peuvent être minimisés avec la création d'une base de départ et l'utilisation d'un menu déroulant pour saisir noms des taxons, noms des auteurs, noms de lieux. Cette approche n'est valable que si la base de départ contient de façon exhaustive l'ensemble des choix possibles. Il est donc nécessaire d'associer un système de contrôle (genre *audit*) de la qualité du processus de numérisation de la donnée, qui ne peut prendre la forme que d'une correction manuelle, ou semi-automatisée, *a posteriori*.

Un problème similaire apparaît lors de la transcription, la translittération ou la romanisation des noms des localités, des régions, des noms des auteurs et des notes écrits dans les langues d'origines. Une des solutions retenues lors des débuts de l'informatisation consistait à n'utiliser que les symboles contenus dans le jeu ASCII correspondant à « l'anglais de base », excluant de fait tous les accents, symboles et diacritiques des autres langues basées sur l'alphabet romain, ainsi que toutes les autres langues n'utilisant pas l'alphabet romain. Toutes les langues ne disposent pas d'un système de translittération permettant de les transcrire en n'utilisant que les caractères latins non accentués. Cette limitation liée au jeu de caractères utilisé conduit à une perte d'information ; une localité parfaitement compréhensible dans sa graphie originale peut devenir ininterprétable quand elle est retranscrite dans un jeu de caractères plus restrictif. Seul un retour à l'étiquette originale permet alors de comprendre la localité ou le nom du collecteur. L'avènement de l'UTF-8 (*Universal Character Set Transformation Format – 8 bits*) comme standard de codage de caractères compatible avec quasiment toutes les langues du monde a permis de retranscrire fidèlement la graphie originale dans les bases, et ainsi toutes les données

6 www.lesherbonautes.mnhn.fr.

des étiquettes des spécimens. Ce standard permet surtout de les échanger sans perte d'information entre systèmes de bases de données. Aujourd'hui, une grande majorité des sites web utilisent ce standard. Ainsi, utiliser un système permettant d'encoder les caractères nativement en UTF-8 garantit un encodage et des échanges de données optimaux entre systèmes.

18.3.2. *Structure des informations liées aux spécimens*

Dans une base de données de collection naturaliste, chaque spécimen doit avoir au moins cinq informations : un identifiant, un nom, des informations qui permettent de le placer dans l'espace et le temps, c'est-à-dire une référence du lieu et de la date de collecte, et un nom de collecteur. Ce dernier est très important car il permet, entre autres, de croiser et vérifier des informations de lieu et date, de donner un crédit intellectuel aux observations et de les associer à des sessions d'échantillonnage spécifiques (*sampling events*). Ceci permet, par exemple, une évaluation de l'effort d'échantillonnage (Caesar *et al.* 2017). Le niveau de précision et les informations additionnelles sur les étiquettes peuvent varier selon la pratique courante d'une période ou d'un collecteur, l'objectif de l'étude qui a généré la collecte, et les informations importantes pour chaque groupe taxonomique (par exemple les botanistes renseignent souvent le type de sol, le type de végétation tandis que la méthode de collecte et l'heure d'échantillonnage sont très importants pour les entomologistes). Dans certains cas, elles peuvent être suffisamment complètes pour générer d'autres informations. Par exemple, les coordonnées géographiques d'un lieu de récolte peuvent être déduites de manière fiable à partir de localités convenablement renseignées. Inversement, des coordonnées géographiques convenablement renseignées permettent d'accéder à toutes les informations sur l'environnement de la récolte qui n'étaient pas mentionnées sur l'étiquette ni dans les carnets de récolte.

Le format dans lequel ces informations sont présentées, le niveau de détail, l'association de plusieurs autres informations (par exemple sol, végétation, espèces hôtes) font que les combinaisons possibles de ces informations sur les étiquettes, ou les cahiers de terrain, peuvent se décliner presque à l'infini. Le deuxième challenge de la mise en disposition des données est donc leur structuration, c'est-à-dire l'association de chaque information de l'étiquette à un champ spécifique de la base de données. Cette structuration doit être la plus développée possible et utiliser des champs standards comme le standard Darwin Core.

18.3.3. *Le cadre taxonomique : une information mouvante*

Tandis que le lieu, la date, l'environnement et la méthode de collecte sont des parties inchangeables de l'information relative à un spécimen, le nom est attribué *a posteriori* ; il peut changer avec l'évolution des connaissances ou voire même demeurer identique mais changer de signification (voir le chapitre 2 par Grandcolas). Par exemple, pour les mammifères, le nombre d'espèces connues a changé de 4 629 en 1993, lors de la 2e édition du *Mammal Species of the World* (Wilson et Reeder 1993) à 6 495 en 2018 (Burgin *et al.* 2018). Si de nouvelles espèces ont été découvertes à partir des nouveaux échantillonnages sur le terrain, une partie de ces cas résulte des changements de noms d'une partie des spécimens.

La référence taxonomique aux spécimens de collections n'est possible que s'il existe un ensemble de noms de taxons placés dans une classification, pour lesquels il y a un accord général au sein de la communauté scientifique. Un premier grand effort dans ce sens est la création des référentiels taxonomiques (par exemple *Catalogue of Life*[7], TAXREF pour la France (Gargominy *et al.* 2018)). Il s'agit de bases de données résultant de mise en commun des différents référentiels qui permettent d'identifier le nom valide d'une espèce, le nom des auteurs, la date de description et une liste exhaustive des synonymes. C'est une condition indispensable pour la constitution de bases de données afin de savoir à quel taxon une donnée se réfère. Malheureusement, les référentiels taxonomiques ne précisent pas toujours explicitement les critères adoptés pour résoudre les contradictions taxonomiques et présentent le plus souvent une seule classification. Certains systèmes comme Botalista[8], développé par les Conservatoire et jardin botaniques de la Ville de Genève), permettent de prendre en compte plusieurs classifications concurrentes. D'autres approches comme la *World Flora Online*[9] se basent sur une classification consensus mise au point et gérée par des groupes d'experts taxonomiques qui se mettent d'accord sur la classification à utiliser. Il y a encore un très long chemin avant d'avoir une liste à jour des noms valides des espèces et surtout une identification à jour des spécimens de collection.

18.3.4. *L'importance du traçage de l'origine des données*

L'importance fondamentale des données de collections pour la recherche scientifique vient du fait que la donnée reste reliée à un spécimen auquel nous pouvons toujours revenir pour vérifier, ajouter ou confronter des informations. La

7 www.catalogueoflife.org.

8 www.botalista.community.

9 www.worldfloraonline.org.

première information indispensable lors de la constitution d'une base de données est donc l'attribution d'un identifiant unique pour chaque spécimen, auquel les informations peuvent être associées et corrigées au cours du temps. La pratique de la plupart des institutions consiste à attribuer des numéros d'identification de 1 à *N* pour chaque ensemble de collection (mammifères, chaque groupe d'insectes, plantes, etc.). Lors de la diffusion de ces données *via* la plateforme GBIF, elles reçoivent un autre numéro unique d'identification (le numéro d'origine est conservé dans un champ à part), et de même pour les listes de séquences nucléotidiques GenBank[10]. Cette pratique universelle dans la communauté des gestionnaires de bases de données permet d'éviter l'incorporation de spécimens en doublons, s'ils étaient mis à disposition de manière répétée par le fournisseur de données *via* différents pipelines par exemple. Mais il faut toujours que ce numéro GBIF revienne à la collection, sans quoi aucune évolution en libre accès de la collection n'est possible. Pour cette raison, la création d'un identifiant unique pour chaque spécimen de collection est un des grands sujets de discussion au sein des infrastructures pour la mise à disposition des données de collection (comme l'infrastructure pan-européenne DiSSCo[11] – The Distributed System of Scientific Collections – ou iDigBio[12] aux États-Unis).

18.4. Mettre à disposition les données : le rôle des portails d'accès

Les différentes institutions éditrices de données partagent leurs différentes bases de données locales *via* des portails d'accès aux données comme le GBIF. Le rôle de ces portails est de permettre à l'utilisateur de retrouver et télécharger toutes les données nécessaires à son travail à partir d'un seul point d'accès. Cependant, nous avons vu que les pratiques de saisie et de préparation des données dans chacune de ces institutions peuvent être très hétérogènes. Un travail de standardisation et d'interprétation est alors opéré lors de la diffusion des données des institutions jusqu'au portail GBIF, une opération supervisée par les points nodaux du GBIF (c'est-à-dire les équipes du GBIF dans chaque pays ; par exemple, le GBIF France).

18.4.1. Les standards de mise à disposition

Le secrétariat du GBIF a développé pour la diffusion des données des différentes institutions une plate-forme logicielle appelée IPT pour *Integrated Publishing Toolkit*. C'est un logiciel libre permettant de transcrire les bases de données selon le

10 www.ncbi.nlm.nih.gov/genbank.

11 www.dissco.eu.

12 www.idigbio.org.

standard Darwin Core développé par le TDWG (Taxonomic Databases Working Group) (Wieczorek *et al.* 2012), un ensemble de normes rendant les bases de données interopérables. Une des premières étapes de l'IPT, appelée *mapping*, consiste à faire correspondre les champs de la base de données des institutions avec ceux du Darwin Core. Par exemple, le champ (ou la colonne d'un tableau) contenant le nom d'espèce doit être identifié en tant que tel (*scientificName* dans le standard Darwin Core) même s'il est nommé autrement dans le tableau d'origine (par exemple *species*, *species-name*, *species_name*, etc.). Cette configuration est mise en place pour chaque nouvelle base de données, puis automatisée pour leurs mises à jour successives. Cinq champs sont obligatoires (*basisOfRecord*, *occurrenceID*, *scientificName*, *eventDate* et *countryCode*), dix autres sont recommandés, et jusqu'à 165 champs peuvent être remplis.

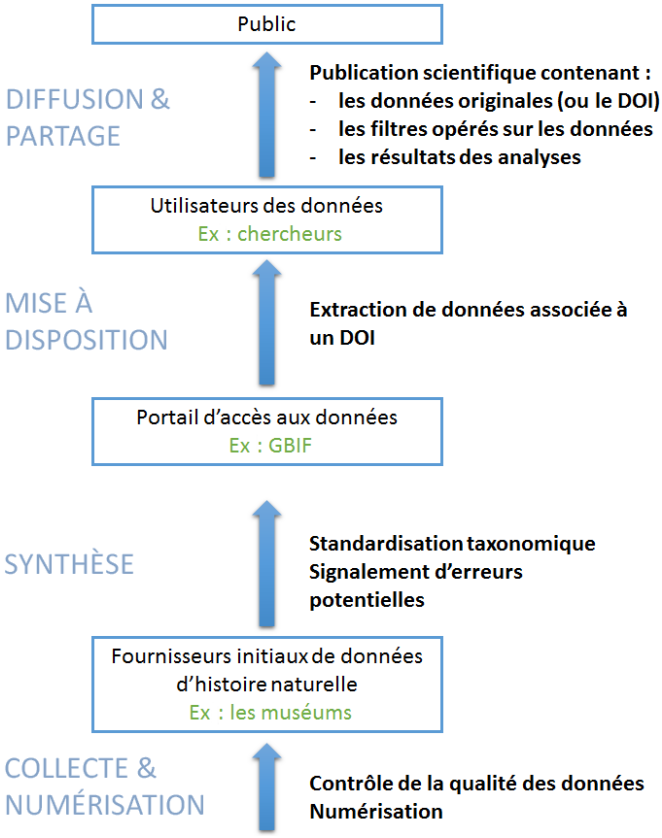

Figure 18.2. *Le chemin des données de leur collecte jusqu'à la diffusion de résultats scientifiques : des processus de vérifications et de standardisation entrepris par différents acteurs. L'utilisateur a la responsabilité finale d'assurer la reproductibilité*

des ultimes opérations effectuées sur les données. Le DOI est un numéro d'identifiant unique qui peut être associé à un export de données.

Ce sont les équipes des points nodaux qui sont en charge de guider les institutions dans cette démarche. En général, les institutions de grande taille hébergent elles-mêmes localement ce logiciel sur leurs serveurs. Les données et métadonnées (c'est-à-dire toutes informations associées au jeu de données dans sa globalité) restent la propriété et sous la responsabilité de leur fournisseur (notamment en matière de qualité des données). Le rôle du GBIF est de fournir une assistance technique aux chercheurs et/ou aux responsables des collections souhaitant mettre à disposition leurs données[13]. Au MNHN par exemple, les données des différentes bases locales (par exemple Jacim, Sonnerat) sont partagées sur le GBIF au fur et à mesure de leurs mises à jour. Lorsqu'une nouvelle institution veut publier des données, le point nodal du pays concerné est averti afin de lui permettre d'assurer et d'offrir un suivi aux fournisseurs.

13 www.gbif.fr/page/contrib/connecter-des-donnees-au-gbif.

Lors de cette étape de diffusion des données des bases institutionnelles *via* le GBIF, un important travail de vérification est également réalisé. Les noms scientifiques sont comparés à ceux de la GBIF Backbone Taxonomy qui sert de liste taxonomique de référence. Elle se base sur le *Catalogue of Life*[14] ainsi que sur de nombreuses autres listes de référence taxonomiques thématiques et nationales. Ils peuvent soit être validés, soit signalés comme erronés et un renvoi à un nom accepté est alors suggéré. Les informations géographiques sont vérifiées et converties vers le format « degrés décimaux » utilisé par le GBIF. Ainsi, de nombreuses erreurs ou imprécisions bien connues sont signalées (*flagged*, en anglais) par les portails d'accès aux données comme le GBIF qui fournissent ainsi une aide non négligeable aux utilisateurs. C'est le cas par exemple lorsqu'une information de latitude ou de longitude est manquante, ou inversée, ou que les coordonnées géographiques sont égales à zéro. Un fournisseur de données a donc tout intérêt à acquérir les données directement en standard Darwin Core (avec un jeu de caractère universel comme UTF-8) pour s'assurer que les informations seront correctement interprétées ultérieurement.

18.5. L'importance du design des analyses scientifiques pour s'approprier les spécificités des données issues des collections

Les données de collection constituent souvent un apport considérable par rapport à d'autres jeux de données (Lavoie 2013). En permettant notamment d'accéder à une dimension historique (Boakes *et al.* 2010), elles documentent les aires de répartition des organismes avant les transformations des paysages par l'homme moderne. Grâce à des données historiques issues des collections d'histoire naturelles, Tóth et collaborateurs (Tóth *et al.* 2014) ont par exemple montré un plus faible *turnover* d'espèces de mammifères actuellement dans les aires protégées au Kenya qu'au début de XX^{e} siècle. Par ailleurs, lorsqu'elles sont comparées à des relevés de végétation standardisés, les données issues d'herbiers fournissent des listes d'espèces plus complètes grâce à la recherche active d'espèces rares par les collecteurs (Garcillan et Eczurra 2011). Les données issues des muséums peuvent aussi avoir une plus large couverture spatiale que d'autres sources de données (par exemple données d'atlas ou données de baguage d'oiseaux focalisées sur l'Europe de l'Ouest (Boakes *et al.* 2010)). Elles représentent une opportunité considérable pour certains groupes comme les invertébrés, souvent négligés par ailleurs (Ponder *et al.* 2001).

14 www.catalogueoflife.org.

Dans la plupart des cas, les données de collections naturalistes ont été acquises afin de constituer une ressource pour la taxonomie et systématique (voir chapitre 2 par Grandcolas et chapitre 20 par Rouchon). Leur utilisation pour des questions de macroécologie ou de macroévolution, plus récente, s'est accompagnée du développement d'un arsenal méthodologique pour faire face aux incertitudes inhérentes à l'acquisition composite des données. De manière générale, il faut veiller à ce que la question posée puisse trouver une réponse grâce à une analyse appropriée des données de collections malgré leur caractère parfois hétéroclite.

18.5.1. Connaître les biais propres aux données des collections : avantages et opportunités pour les analyses scientifiques

Les opportunités qu'offrent les données des collections ne doivent pas faire oublier leurs limites et spécificités. La mise à disposition de ces données reste souvent bien loin d'être achevée. Il est vrai que certains groupes taxonomiques sont mieux étudiés et échantillonnés que d'autres, ce qui produit des différences entre le nombre d'espèces représentées dans les collections et leur diversité réelle. Ceci est très visible, par exemple, quand on compare l'exhaustivité de représentation dans les bases entre vertébrés *versus* non-vertébrés (Troudet *et al.* 2017), ou encore entre papillons et grands coléoptères *versus* des petites mouches (Ponder *et al.* 2001). En outre, les possibilités de numérisation sont également à l'origine d'énormes disparités dans la mise à disposition des données. Par exemple dans les grandes collections, les plantes sont presque totalement numérisées (LeBras *et al.* 2017) grâce à la numérisation en chaîne des parts d'herbier. À l'autre extrême, la numérisation de la plupart des groupes d'insectes avance à pas très lents du fait de la fragilité des spécimens, de leur conservation en 3D et des besoins de manipulation des étiquettes.

Les biais géographiques ont été souvent remarqués dans des données des collections. Les collectes à proximité des routes, près des centres universitaires et dans les endroits les plus accessibles (Pautasso et McKinney 2007 ; Daru *et al.* 2018) tendent à être plus fréquentes et complètes. Selon l'échelle géographique considérée, d'autres facteurs, comme les différences économiques et académiques entre pays, jouent aussi un important rôle (Araújo 2003). À l'échelle globale, les différences d'effort d'échantillonnage couplées avec la capacité des pays riches à mettre rapidement leurs données à disposition mènent à des distorsions importantes de l'image de la distribution de la biodiversité sur la planète. Par exemple, l'image de l'ensemble des occurrences accessible *via* le GBIF montre que la quantité de données disponibles pour les pays les plus développés, et avec forte tradition

académique, est plusieurs fois plus importante que celle de l'Amazonie ou du bassin du Congo, les endroits avec la plus grande diversité biologique de la planète.

Malgré la grande valeur des données historiques pour des études diverses et pour les suivis des changements de la biodiversité à travers le temps (chapitre 16 par Muller *et al.*), peu d'études se sont intéressées à évaluer leurs biais temporels. Néanmoins, l'étude de Daru *et al.* (2018) indique que celles-ci sont aussi biaisées temporellement, suivant les pics d'activité des grands collecteurs et des moments historiques importants dans les trois pays examinés. Boakes *et al.* (2010) ont montré une réduction du nombre d'espèces récoltées entre 1930-1940, la période de la Grande Guerre en Europe.

18.5.2. Vers une bonne adéquation entre la question et les données disponibles

Comme dans tout projet de recherche, le grand défi intellectuel est de bien formuler la question à l'étude et de mesurer la disponibilité et la pertinence des données qui peuvent permettre de répondre à cette question. Il faut ensuite s'approprier ces données en les ré-échantillonnant et en sélectionnant la fenêtre spatio-temporelle adéquate. Il est également nécessaire d'identifier les ressources auxiliaires disponibles qui peuvent faciliter le contrôle des données et corriger des biais éventuels.

On peut donc envisager plusieurs situations, depuis les questions pour lesquelles les données ne seront pas disponibles ou pertinentes ou celles plus positives pour lesquelles des validations en quantité et qualité raisonnables et des traitements statistiques appropriés permettront de ré-échantillonner les données et d'obtenir des éléments de réponse même limités.

Un bon exemple d'une situation positive est fourni par Delisle *et al.* (2003) qui ont pu comprendre la dynamique d'invasion de plantes invasives dans le sud du Québec à partir de données historiques d'herbier. Comment sont-ils parvenus à surmonter les biais inhérents aux processus de collecte ? Les plantes n'avaient alors pas été récoltées dans le but de mesurer l'expansion de leur répartition. L'effort d'échantillonnage n'est pas uniforme au cours du temps ni dans l'espace, et donc l'augmentation de la collecte d'espèces exotiques dans les collections peut à la fois présager d'un phénomène d'expansion tout comme d'une augmentation de l'effort d'échantillonnage. Les auteurs ont corrigé ce biais grâce à l'étude conjointe de l'échantillonnage des espèces natives collectées simultanément : ils ont comparé l'espace cumulé au cours du temps occupé par chaque espèce exotique à celui occupé par les espèces natives. Lorsque la proportion (d'espèces exotiques par

rapport aux espèces natives) augmente pendant une période donnée, cela suggère fortement que la superficie occupée par les espèces exotiques augmente réellement, parce qu'elle augmente plus rapidement que si elle était strictement le résultat d'une meilleure couverture spatiale de l'échantillonnage des spécimens d'herbiers (Delisle *et al.* 2003).

18.5.3. *Jouer l'atout des échelles spatiales multiples*

Les localités géographiques associées aux spécimens de collection représentent une énorme quantité d'information, mais qui comporte de grandes disparités en matière de précision, notamment avec les données historiques dont la précision est souvent faible. Ces données sont souvent écartées des analyses statistiques qui peuvent exiger des données d'occurrences à haute précision acquises avec des GPS. Pourtant ces données, malgré leur faible précision, apportent une information qui peut être essentielle. Les chercheurs se sont saisis de cet enjeu et ont construit des modèles d'estimation des distributions spatiales qui se nourrissent à la fois de données à fine et faible précision (Keil *et al.* 2014). La synergie entre les différents types de données s'avère souvent cruciale du fait de leur complémentarité, les données historiques permettant de contrôler par exemple la complétude des répartitions estimées avec des données récentes limitées mais précises.

18.6. Le passage des données brutes à des données triées utilisables pour les analyses scientifiques

La mise en commun et l'augmentation de la taille des jeux de données a entraîné une augmentation de la complexité des jeux de données et des analyses. De nouvelles compétences s'avèrent donc nécessaires pour construire et assembler ces bases de données, mais aussi pour les analyser.

La première étape d'une analyse scientifique, une fois la question bien délimitée, consiste donc à détecter les erreurs résiduelles dans les jeux de données constitués et à les écarter ou les corriger (figure 18.3). Cependant, du fait du grand nombre de données devenues disponibles durant ces dernières décennies, chaque donnée ne peut généralement pas être vérifiée et corrigée individuellement. Les erreurs sont inévitables dans un vaste ensemble de données. Elles sont de nature et de sources multiples et leur présence est d'autant plus probable que les sources de données d'une même base sont multiples et que la base est grande.

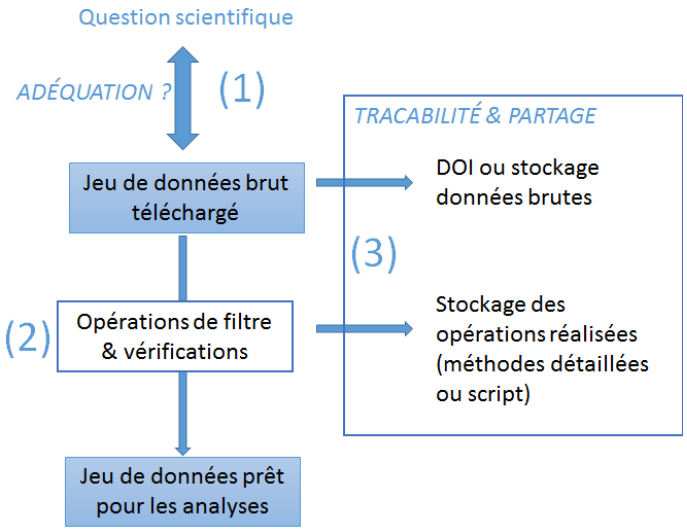

Figure 18.3. *Des données téléchargées aux données préparées pour l'analyse : 1) vérification de l'adéquation des données avec la question posée, 2) opérations sur les données et 3) responsabilité de l'utilisateur d'assurer la traçabilité des données utilisées et des opérations réalisées.*

Le premier type d'erreurs se trouve dans les informations taxonomiques : les noms attribués aux taxons peuvent avoir été mal orthographiés ou des synonymes avoir été utilisés à la place des noms valides. Ces erreurs sont en général filtrées dès l'étape de l'informatisation grâce aux méthodes de standardisation développées par les fournisseurs de données. Cependant, certaines passent entre les mailles de ces filets. En 2016, Zermoglio et collaborateurs (Zermoglio *et al.* 2016) ont analysé les erreurs retrouvées dans VertNet, un agrégateur de données réunissant plusieurs centaines de collections de vertébrés[15]. Ils ont montré que seulement 47 % des taxons présentaient des noms valides mais que 97 % des taxons peuvent être associés à un nom valide quand les noms contenus dans la base de données sont confrontés à un référentiel. La solution consiste ainsi à comparer la liste de noms de taxons extraite à un référentiel taxonomique. Pouvoir se référer à un référentiel à jour qui fait autorité est donc fondamental. Des outils techniques existent pour automatiser la mise en correspondance d'un nom de taxon avec le nom accepté

15 www.vertnet.org.

(voir, par exemple, (Cayuela *et al.* 2012), ou le webservice Biodatascreen produit par notre équipe de recherche[16]).

Le second type d'erreurs rencontrées fréquemment sont les erreurs inhérentes aux données géo-spatialisées. Ce sont des erreurs récurrentes mais en général bien connues. En 2019, Zizka et collaborateurs (Zizka *et al.* 2019) ont détecté, parmi les 90 millions d'occurrences de plantes à fleurs, 3,4 millions d'occurrences potentiellement problématiques (3,7 %). Là encore, des solutions d'automatisation pour la détection des erreurs existent et sont souvent signalées par les fournisseurs de données. Cependant, s'il est facile de détecter et d'écarter des coordonnées de géolocalisation d'une plante terrestre qui tomberaient dans l'océan, ou des coordonnées douteuses (par exemple avec une latitude et/ou une longitude égale(s) à zéro, ou des coordonnées non converties en degrés décimaux), il est bien plus compliqué de détecter des coordonnées qui semblent normales, mais qui tombent hors de l'aire connue et validée d'une espèce. Il est alors essentiel de pouvoir distinguer une identification erronée d'une station valide, par la validation d'un expert ou la confrontation des données avec un jeu de données indépendant. Certaines bases de données fournissent des données de présence à l'échelle sub-nationale ou nationale qui s'avèrent très utiles pour valider des données d'occurrences. Par exemple, Veron *et al.* (2019) ont utilisé la base *e-monocot*[17] pour aider à l'exclusion des occurrences non natives des espèces dans leurs analyses. De même, Monnet *et al.* (2021) ont écarté des occurrences douteuses d'arbres méditerranéens en confrontant les données à l'inventaire d'espèces le plus récent mené à l'échelle des pays et de leurs îles (Médail *et al.* 2019).

18.6.1. *De l'*open data *à l'*open science*, une responsabilité sur la traçabilité des données et des opérations réalisées*

L'utilisateur des données doit ainsi être capable de détecter les erreurs résiduelles pour obtenir un jeu de données de travail dont il est capable d'évaluer la quantité et qualité (figure 18.2, remarque « Quelques bonnes pratiques pour l'utilisateur de données »). Cependant, comme l'utilisateur n'a pas la main pour modifier les données brutes, la plupart du temps, les données avec erreurs ou des informations incomplètes sont seulement écartées de l'analyse.

Il est alors fondamental de pouvoir conserver les données ré-échantillonnées et les changements opérés pour de futures analyses ou ré-analyses. Ceci nécessite en général d'enregistrer tous les traitements réalisés sur les données par codage en

[16] www.bloom.snv.jussieu.fr/biodatascreen/.

17 www.about.e-monocot.org.

utilisant un langage de programmation. Ces filtres doivent être transparents pour améliorer la reproductibilité des analyses. De plus, pour pouvoir répliquer ces analyses, un nouvel utilisateur potentiel doit avoir accès au même export de la base de données sur laquelle les premiers utilisateurs ont travaillé. Ainsi, une condition à la reproductibilité des analyses implique de communiquer le DOI associé au jeu de données fourni par certains fournisseurs de données (comme le fait GBIF), ou à défaut de stocker la base de données dans des archives durables. Dans plusieurs décennies, un chercheur devrait être capable d'accéder à la même base de données utilisée pour les analyses réalisées aujourd'hui.

L'identifiant DOI fourni par le GBIF pour un jeu de données permet également de relier les données qui ont été utilisées à une publication : lors de la consultation de données sur GBIF, un utilisateur peut voir comment celles-ci ont été préalablement utilisées. L'enjeu est aussi de rendre le travail plus collaboratif : l'effort mis pour comprendre les données et les nettoyer n'est plus à faire et peut être partagé et amélioré. L'*open data* (les données ouvertes) nous porte ainsi rapidement vers l'*open science* (la science ouverte) avec un partage croissant des analyses en plus de celui des données. De plus en plus d'outils sont disponibles pour le traitement de données, ils se démocratisent et gagnent en efficacité. Le langage de programmation SQL (*Structure Query Language*) dédié à la manipulation des bases de données relationnelles et le langage de programmation R (R Core Team 2019) sont largement utilisés par les scientifiques. Récemment, le logiciel OpenRefine[18], un outil conçu « pour travailler sur données en désordre », permet de faire un pas vers l'automatisation pour les personnes non familières avec la programmation pour les étapes de détection et de correction des différences orthographiques dans les bases de données. En effet, ce logiciel permet d'opérer manuellement des modifications d'un jeu de données sans formules tout en conservant une trace de ces dernières, offrant ainsi la possibilité de répéter les opérations effectuées une première fois.

Les corrections opérées par l'utilisateur ne sont donc pas apportées directement sur les données brutes, mais seulement *enregistrées*. En effet, dans de nombreux cas l'utilisateur des données n'est pas le même que le fournisseur. Ceci montre l'intérêt d'instaurer un lien fournisseur/utilisateur afin de mettre en place une boucle de rétroaction positive, de l'utilisateur au fournisseur des données, qui pourra accélérer la correction des erreurs à la source.

18 www.openrefine.org.

18.6.2. Vers une nécessaire réorganisation du travail collaboratif

Un enjeu supplémentaire réside dans le fait que pour les besoins d'une analyse scientifique, des données de natures différentes sont souvent mobilisées, chacune avec leur lot d'erreurs à détecter, et à écarter ou à corriger. Ainsi, une même question scientifique peut nécessiter de rassembler des données systématiques d'occurrence, génétiques, phylogénétiques, environnementales, etc. Les questions de recherches requièrent donc une interopérabilité entre différents types de données et de leurs outils de validation. L'analyse scientifique doit pouvoir s'appuyer sur une séquence de travail bien structurée (*workflow*, en anglais).

Cette synthèse peut nécessiter également de nouveaux outils de collaboration pour permettre l'interopérabilité entre les différentes cellules d'expertises, notamment quand celles-ci sont gérées par des personnes différentes. Par exemple, le projet FunctionalWebs du Centre de synthèse et d'analyse sur la biodiversité (CESAB, France) a mis en place un *package* sur le logiciel R[19] offrant à chaque collaborateur un accès à la base de données en constante évolution. Ainsi, tout en poursuivant la collecte et la mise à jour des données, les différents membres de leur consortium pouvaient d'ores et déjà travailler en simultané sur la dernière version mise à jour de la base de données.

REMARQUE. Quelques bonnes pratiques pour l'utilisateur de données

Idéalement, tout traitement des données, pour être reproductible, doit être fait sous la forme d'un script (R ou autre langage) et inclure toutes les modifications apportées aux données brutes. Ci-après sont résumées les bonnes pratiques à chaque grande étape de traitement et de vérification des données.

– Dès le téléchargement des données : des informations seront nécessaires pour la publication des résultats associés aux données et peuvent être récoltées en amont :

- stocker une version brute de l'export réalisé pour archivage ;

- prendre note de la manière de citer le jeu de données (par exemple la citation et la date d'accès aux données) ;

- prendre note du DOI si disponible.

– Design du projet scientifique – vérifier la pertinence des données mobilisées pour la question : les questions suivantes peuvent être posées pour interroger la base de données et vérifier l'adéquation entre les données et la question scientifique posée :

19 www.github.com/SrivastavaLab/fwdata.

- combien de données sont fournies ? Pour combien de taxons ?

- combien de coordonnées uniques ?

- tous les champs nécessaires à l'analyses sont-ils disponibles ? Combien de données sont manquantes dans chaque champ requis ?

- des données ont-elles été signalées comme douteuses par le fournisseur ? Si oui, combien ?

Ces questions préliminaires permettent d'évaluer si les données répondent aux attendus quant à la question scientifique posée. Dans le cas contraire, des données complémentaires peuvent être mobilisées et/ou les contours de la question doivent être redéfinis (dans sa dimension taxonomique, spatiale et/ou temporelle).

– Les opérations de filtres des données : sous-échantillonner les données pour se les réapproprier :

- les noms de taxons sont-ils des noms acceptés ou nécessitent-ils une mise en correspondance avec le nom accepté (synonymie) selon un référentiel taxonomique en vigueur (standardisation taxonomique selon un référentiel taxonomique) ?

- les erreurs les plus récurrentes dans les données géolocalisées (Zizka *et al*. 2019) ont-elles été identifiées préalablement par le fournisseur de données ou doivent-elles être détectées par l'utilisateur (élimination des données douteuses identifiées par l'utilisateur et/ou le fournisseur) ?

- des erreurs supplémentaires peuvent-elles être détectées en visualisant la distribution des données pour chaque taxon (ou groupe supérieur) ou en confrontant les données obtenues à un jeu de données indépendants ou un avis d'expert (élimination des potentielles erreurs d'identification) ?

– La publication :

- les différentes bases de données sont-elles correctement citées et accessibles *directement* au public, soit sur un répertoire d'archivage, soit *via* un identifiant DOI fourni par la base de données source ?

- les tris réalisés sur les données sont-ils reproductibles grâce à des méthodes suffisamment décrites ou l'archivage du script ?

18.7. Conclusion

Les études utilisant des spécimens et des données issues des collections d'histoire naturelle se sont multipliées depuis le début des années 1990 (Lavoie 2013 ; Suarez et Tsuitsui 2013). Elles traitent un très grand nombre de sujets de recherche différents et mobilisent un nombre croissant de spécimens informatisés ou numérisés. Ces données emblématiques de la science ouverte permettent à la fois un accès rapide à l'univers de la biodiversité et un changement d'échelle taxonomique, géographique ou temporel. Grâce à elles, des questions cruciales pour la science ou la société peuvent être posées, qui embrassent une grande partie du Vivant, du globe terrestre ou s'appuient sur des points de référence (« points zéro ») pour définir des tendances. L'ensemble de ces questions serait autrement impossible à traiter à moins d'investissements colossaux ou d'attentes prolongées pour acquérir de médiocres équivalents de jeux de données.

Utiliser ces données nécessite cependant une expertise particulière ; il faut les mobiliser, les assembler, les valider et les ré-échantillonner et ces étapes représentent un travail important et sophistiqué. Cependant, cette expertise se développe dans la communauté scientifique et des méthodes et des outils apparaissent et se démocratisent, permettant aux utilisateurs d'opérer leur propre tri des données.

Ce faisant, nous bénéficions de données ouvertes et il nous appartient d'utiliser ces données et de réaliser nos analyses avec la même éthique, en veillant à la reproductibilité des opérations effectuées. La mise en commun des données (*données ouvertes*) nous porte ainsi rapidement vers la *science ouverte*, encourageant un partage croissant des analyses en plus de celui des données. Réinsérer les collections d'histoire naturelle dans les problématiques scientifiques et sociétales implique la synthèse de jeux de données différents mais également la synthèse d'expertises et de connaissances différentes, appelant à repenser l'organisation du travail d'analyses scientifiques vers plus de collectif. Les spécificités des données de collection (par exemple l'existence de données historiques) représentent des opportunités incroyables pour ouvrir le champ des connaissances, mais les biais associés ne doivent pas être négligés dans les analyses. Au contraire, une meilleure connaissance de ces biais peut guider les choix de la résolution spatiale et temporelle des analyses qui peuvent être menées ainsi que la préparation des futures expéditions pour la collecte de nouvelles données.

Dans cette perspective, plusieurs aspects méthodologiques sont en plein renouvellement et évolution au niveau mondial. Les référentiels taxonomiques sont en construction et en évolution permanente et doivent rendre compte de l'état des

connaissances sur la biodiversité. Il se basent sur des milliers de travaux concernant plus de deux millions d'espèces d'organismes connues. Ce progrès ne va pas sans que des identifiants uniques soient définis et utilisés pour chaque spécimen conservé, ou pour chaque jeu de données mobilisé. Ce sont là des enjeux qui peuvent paraître simples mais demandent un consensus et une organisation sans faille au niveau international.

Mais le plus grand progrès nécessaire concerne la prise de conscience que les données de biodiversité – y compris celles que nous récoltons actuellement – doivent rester ouvertes et utilisables pendant de longues années. La condition *sine qua non* pour une telle disponibilité est la référence à un échantillon matériel ou *a minima* à des données numériques riches (Grandcolas 2017, 2019 ; Troudet *et al.* 2018). Seule cette référence permet de pallier les imprécisions, les erreurs de travaux qui ont généré les données de biodiversité, mais aussi l'évolution des taxonomies et des noms. La disponibilité des données ne va pas sans la disponibilité des analyses et des étapes intermédiaires de traitement de données qui – elles aussi – doivent être disponibles pour référence ultérieure.

En ce moment de crise de la biodiversité, la communauté scientifique a sans doute trop tendance à découvrir le trésor et le patrimoine que représentent les collections d'histoire naturelle et les données de biodiversité qui y sont résidentes, à l'utiliser avec brio mais en oubliant que nous devons non seulement tirer des « profits » scientifiques de ce patrimoine, mais aussi le conserver et le faire fructifier pour les générations futures.

18.8. Bibliographie

Araújo, M.B. (2003). The Coincidence of People and Biodiversity in Europe. *Global Ecology and Biogeography*, 12(1), 5–12.

Boakes, E.H. *et al*. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol*., 8, e1000385.

Brummitt, R.K. (2001). World Geographical Scheme for Recording Plant Distributions, 2e édition. Document, Biodiversity Information Standards (TDWG).

Burgin, C.J., Colella, J.P., Kahn, P.L., Upham, N.S. (2018). How many species of mammals are there?. *J. Mammal*., 99, 1–14.

Caesar, M., Grandcolas, P., Pellens, R. (2017). Outstanding Micro-Endemism in New Caledonia: More than One out of Ten Animal Species Have a Very Restricted Distribution Range. *PLoS ONE*, 12(7).

Cayuela, L. *et al.* (2012). Taxonstand: An r package for species names standardisation in vegetation databases. *Methods Ecol. Evol*., 3, 1078–1083.

Daru, B.H. *et al.* (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol.*, 217, 939–955.

Delisle, F. *et al.* (2003). Reconstructing the spread of invasive plants: Taking into account biases associated with herbarium specimens. *J. Biogeogr.*, 30, 1033–1042.

Garcillán, P.P., Ezcurra, E. (2011). Sampling procedures and species estimation: Testing the effectiveness of herbarium data against vegetation sampling in an oceanic island. *J. Veg. Sci.*, 22, 273–280.

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Dupont, P., Daszkiewicz, P., Poncet, L. (2018). TAXREF v12, référentiel taxonomique pour la France : méthodologie, mise en oeuvre et diffusion. Rapport, Muséum national d'Histoire naturelle, Paris.

Grandcolas, P. (2017). Loosing the connection between the observation and the specimen: a by-product of the digital era or a trend inherited from general biology?. *Bionomina*, 12, 57–62.

Grandcolas, P. (2019). The Rise of "Digital Biology": We need not only open, FAIR but also sustainable data!. *Biodiversity Information Science and Standards*, 3, e37508.

Haevermans, T. *et al.* (en préparation). A majority of plant life globally at risk as predicted by machine learning threat analysis.

Keil, P. *et al.* (2013). Downscaling of species distribution models: a hierarchical approach (R Freckleton, Ed.). *Methods Ecol. Evol.*, 4, 82–94.

Lavoie, C. (2013). Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspect. Plant Ecol. Evol. Syst.*, 15, 68–76.

Le Bras, G. *et al.* (2017). The French Muséum National d'histoire Naturelle Vascular Plant Herbarium Collection Dataset. *Scientific Data*, 4, 1–16.

Médail, F. *et al.* (2019) What is a tree in the Mediterranean Basin hotspot? A critical analysis. *Forest Ecosystems*, 6 (17).

Monnet, A.-C. *et al.* ((2021). WOODIV, a database of occurrences, functional traits, and phylogenetic data for all Euro-Mediterranean trees. *Sci Data*, 8, (89).

Pautasso, M., McKinney, M.L. (2007). The Botanist Effect Revisited: Plant Species Richness, County Area, and Human Population Size in the United States. *Conservation Biology*, 21(5), 1333–40.

Ponder, W.F. *et al*. (2001). Evaluation of museum collection data for use in biodiversity assessment. *Conserv. Biol.*, 15, 648–657.

R Core Team (2019). R: A language and environment for statistical computing [En ligne]. R Foundation for Statistical Computing, Vienne. Disponible à l'adresse : https://www.R-project.org/.

Rouhan, G. *et al.* (2014). The herbonauts website: recruiting the general public to acquire the data from herbarium labels. Dans *Botanists of the twenty first century: roles, challenges and opportunities*. UNESCO, Paris.

Suarez, A.V., Tsutsui, N.D. (2004). The Value of Museum Collections for Research and Society. *Bioscience*, 54, 66–74.

Tóth, A.B. *et al*. (2014). A century of change in Kenya's mammal communities: Increased richness and decreased uniqueness in six protected areas. *PLoS One*, 9(4): e93092.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, 7(1), 1–14.

Troudet, J., Vignes-Lebbe, R., Grandcolas, P., Legendre, F. (2018). The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it?. *Systematic biology*, 67(6), 1110–1119.

Veron, S. *et al.* (2019). Vulnerability to Climate Change of Islands Worldwide and Its Impact on the Tree of Life. *Scientific Reports*, 9(1).

Wieczorek, J. *et al*. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PLoS One*, 7(1), e29715.

Wilson, D.E., Reeder, D.M. (1993). *Mammal Species of the World: A Taxonomic and Geographic Reference*, 2e édition. Smithsonian Institution Press, Washington.

Zermoglio, P.F. *et al.* (2016). A standardized reference data set for vertebrate taxon name resolution. *PLoS One*, 11, 1–20.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C. *et al.* (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.*, 10, 744–751.