Diplôme national de master

Domaine - sciences humaines et sociales

Mention – sciences de l'information et des bibliothèques

Parcours – archives numériques

Mémoire de recherche / septembre 2023

Documenter la biodiversité, entre données numériques et spécimens physiques

Mélusine Rocher

Sous la direction de Clément Oury
Conservateur des bibliothèques et Adjoint au chef du service Conservation, Restauration, Numérisation – Muséum national d'Histoire naturelle

Remerciements

Un grand merci à Clément Oury pour m'avoir guidée tout au long de ce sujet de recherche, ainsi que pour ses retours au cours de la rédaction.

À Hugo, Irène, Audrey et mes parents, dans le désordre : ma reconnaissance pour vos relectures et soutiens respectifs, ainsi que pour les heures de travail en commun qui ont été fructueuses. Ce mémoire aurait été bien différent sans vous pour m'y accompagner.

Mes remerciements à l'ensemble de ma famille (et plus particulièrement aux deux petits diablotins adorables que sont mes sœurs), pour son soutien sans faille, les distractions de dernière minute, la bonne humeur contagieuse et pour m'avoir appris la curiosité.

Je souhaiterais également faire une place pour l'ensemble des personnes qui m'ont accompagnée sur ces dernières années et qui m'ont permis de me construire en tant qu'adulte.

Enfin, une pensée toute particulière pour mon « papi Castor », qui m'a, il y a une bonne quinzaine d'années déjà, insufflé une curiosité du monde qui nous entoure. Les mercredis après-midi passés au MNHN dans la Grande Galerie de l'Évolution, aux expositions temporaires ou encore à la Paléontologie du Jardin des Plantes sont certainement, sans que je ne m'en sois rendue compte en septembre dernier, la raison du choix de ce sujet. Même si les liens se sont quelque peu distendus, je garde toujours ces moments hors du temps dans un petit coin de ma tête.

Résumé :

La taxonomie et la systématique sont deux disciplines de la biologie, chargées respectivement de décrire et de classifier le vivant. Toutefois, avec l'arrivée des nouvelles technologies numériques et du web, elles se retrouvent face à de nouvelles pratiques qui modifient les fondamentaux mêmes de leurs façon de faire de la recherche. Nous étudierons le concept de spécimen et son évolution à travers les nouvelles données numériques qui peuvent maintenant être mobilisées par les taxonomistes pour décrire de nouvelles espèces, et, de façon plus générale, nous nous pencherons sur l'archivage des données de la biodiversité.

Descripteurs : taxonomie intégrative – spécimen – spécimen numérique étendu – bio-informatique – bases de données – données de la recherche – archivage

Abstract :

Taxonomy and systematics are two sub-fields of biology, respectively dealing with describing and classifying Earth's biodiversity. However, due to new digital technologies and the Web, they both face next-generation practices challenging the basics in their way to conduct research. We will study the concept of "specimen" and its evolution through new digital data that taxonomists may use to describe new species, and we will examine the archiving and long-term preservation of biological research data

Keywords : integrative taxonomy – specimen – digital extended specimen – bioinformatics – databases – research data – preservation – archiving

Droits d'auteurs

Sommaire

Sigles et abréviations

ABCD : Access to Biological Data Collection
ACEF : Annual Checklist Exchange Format
ADN : acide désoxyribonucléique
ALA : Atlas of Living Australia
API : Application Programming Interface
ASAP : Assemble Species by Automatic Partitioning
BICIKL : BIodiversity Community Integrated Knowledge Library
BioCASe : Biological Collection Access Service
CC : Creatives Commons
CETAF : Consortium of European Taxonomic Facilites
CLS-JSON : Collecte Localisation Satellites – JavaScript Object Notation
CNRS : Centre National de la Recherche Scientifique
COI : Cytochrome *c* Oxidase I
CoL : Catalogue of Life
COLDP : Catalogue of Life Data Package
CSV : Comma-Separated Values
DES : Digital Extended Specimen
DISSCO : Distributed System of Scientific Collections
DOI : Digital Object Identifier
DwC : Darwin Core
EMBL : European Molecular Biology Laboratory
EML : Ecological Metadata Language
ENA : European Nucleotide Archive
EPHE-PSL : École Pratique des Hautes Études – Paris Sciences et Lettres
FAIR : Findable, Accessible, Interoperable, Reusable
GBIF : Global Biodiversity Information Facility
GPS : Global Positioning System
GMYC : Generalized GMyC Mixed yule Coalescent
HTTP : Hypertext Transfer Protocol
IIIF : International Image Interoperability Framework
ISO : International Organization for Standardization
ITIS : Integrated Taxonomic Information System
IUCN : International Union for Conservation of Nature
KML : Keyhole Markup Language
KNB : Knowledge Network for Biocomplexity
JSON : JavaScript Object Notation
METS : Metadata Encoding and Transmission Standard
MNHN : Muséum National d'Histoire Naturelle
NHM : Natural History Museum
OAIS : Open Archival Information System
RDF : Ressource Description Framework
sp. : *species*
SPART : Species PARTition
TDWG : Taxonomic Database Working Group
TSV : Tab Separated Values
UMR : Unité Mobile de Recherche
URI : Unique Ressource Identifier
WoRMS : World Register of Marine Species
XLS : Excel

XML : Extensible Markup Language
YAML : Yet Another Markup Language Ain't Markup Language
3D : 3 Dimensions

INTRODUCTION

Depuis quelques décennies déjà, les scientifiques et les militants écologistes attirent l'attention du grand public sur la disparition de masse du vivant : à la suite des cinq grandes extinctions qu'a connu la Terre au cours de son histoire, on en liste aujourd'hui une sixième, causée en grande partie par l'impact des activités humaines sur Terre. Une étude récente a ainsi estimé qu'entre 7,5 et 13 % des 2 millions d'espèces connues ont disparu au cours des six cents dernières années, en extrapolant à partir des chiffres obtenus pour les mollusques (Cowie et al., 2022). Même si ces chiffres sont peut-être légèrement exagérés (comme le reconnaissent eux-mêmes les auteurs de l'article), ils mettent bien en avant que les espèces disparaissent aujourd'hui à un rythme sans précédent depuis l'arrivée des humains sur Terre.

Dans ce contexte, où la biologie voit disparaître ses objets d'étude, la taxonomie et la systématique se retrouvent face à une situation de plus en plus urgente : ces deux disciplines, chargées respectivement de décrire et de classifier le vivant, donnent en effet aux autres disciplines de la biologie la clef d'identification des espèces pour que celles-ci soient étudiées et puissent, entre autres, faire l'objet de politiques de conservation. L'enjeu de leur travail est alors de conserver une trace relativement pérenne du vivant avant qu'il ne disparaisse, et cette trace se doit d'être suffisamment détaillée pour être réutilisable par les différents biologistes.

Depuis une trentaine d'années, avec la démocratisation des technologies numériques et de leur applications, le terme de *bioinformatics* (ou bio-informatique en français) a fait son apparition au sein de la biologie : il s'agit alors de qualifier l'utilisation du numérique et de l'information pour la biologie, en particulier les procédures de qualification des données pour permettre leur exploitation par les chercheurs (Gadelha et al., 2021). Ces données sont de nature diverse : séquençages ADN, imagerie par rayons X, observations d'animaux relevées par des caméras automatiques, etc. Fortement investies par les autres disciplines de la biologie, ces données sont issues des observations réalisées par les scientifiques eux-mêmes : elles présentent une variété de formats et sont souvent éparpillées dans différentes bases de données sur le Web.

Les taxonomistes et les systématiciens, chargés de décrire et de classifier le vivant, se retrouvent alors pris entre deux feux : d'un côté, la numérisation croissante des pratiques, qui changent les données avec lesquelles ils peuvent travailler, et d'un autre côté, la nécessité de se retrouver face aux animaux et aux plantes qu'ils cherchent à classifier. Le spécimen, c'est-à-dire l'individu conservé dans les collections d'histoire naturelle, est alors le document privilégié de la taxonomie car preuve directe de ce que l'on a collecté et objet auquel on peut se référer pour extraire de nouvelles informations, voire pour corriger la première identification qui a été proposée.

Face à ce changement de situation, les taxonomistes se retrouvent face à différents interlocuteurs, dont certains à l'intérieur même de leur discipline : certains taxonomistes se sont en effet emparés des nouveaux outils, tandis que d'autres n'ont pas pris le tournant numérique et continuent à décrire les espèces comme avant. En dehors de la taxonomie, les autres biologistes produisent des données qui peuvent être mobilisées, en complément d'un spécimen physique, par les taxonomistes pour décrire de nouvelles espèces. Enfin, notons que les institutions naturalistes ont également leur rôle à jouer, dans la mesure où elles hébergent depuis parfois des siècles des collections de spécimens, qui servent aux

taxonomistes. Ces institutions ont également été confrontées au tournant du numérique, avec le passage aux catalogues informatisés et la mise en ligne de leurs collections, mais également avec les différents travaux de recherche scientifiques qu'elles financent.

Avec les collections naturalistes, nous touchons à une autre problématique : celle de la conservation sur le long terme des documents de la taxonomie. Les spécimens présents dans les collections peuvent en effet être mobilisés sur une longue durée, faire l'objet de réattributions scientifiques et servir à de nouveaux trajets de recherche ; mais les données produites par les autres champs de la biologie, si elles sont mobilisées par les taxonomistes, sont à leur tour concernées par l'accessibilité et la réutilisation sur une longue durée.

Avec l'arrivée du numérique et des nouvelles technologies informatiques, comment les documents de la taxonomie ont-ils évolué ? Quelles sont les nouvelles approches ? À quels nouveaux documents mènent-elles ?

Nous avons mobilisé plusieurs approches au cours de ce travail de recherche. Après avoir exploré la littérature scientifique, nous avons réalisé trois entretiens semi-directifs, avec une grille de questions personnalisée pour chacune des personnes interrogées. Le but était alors de confronter nos lectures avec la réalité du terrain, dans la mesure où la littérature scientifique nous paraissait très optimiste sur certaines questions. Nous avons également exploré différentes ressources taxonomiques disponibles en ligne, afin de mieux comprendre la nature des données qui y sont déposées. Confronter les différences de vues nous a également permis d'adopter une posture réflexive sur la situation. Par ailleurs, le présent travail cherche à rendre accessibles à des professionnels de l'information certaines des notions et discussions clefs de la taxonomie telle qu'elle est pratiquée en ce moment, afin de faciliter les discussions entre archivistes, bibliothécaires et taxonomistes.

En prenant ici la taxonomie comme fil rouge de nos réflexions, nous souhaiterions aborder, de façon plus générale, la question des données de la biodiversité, entre numérique natif, numérisations et spécimens physiques. Nous commencerons ainsi par étudier la taxonomie et son contexte, à savoir les différentes manières de faire qui existent au sein de cette science aussi bien que les évolutions de celles-ci, qui amènent à repenser le spécimen. C'est ce que nous étudierons dans un second temps : l'évolution même du concept de spécimen, sa numérisation et son extension, avant de voir son éventuelle mise en place par les collections naturalistes et son utilisation par les chercheurs. Enfin, portés par la question du document et de sa conservation sur le long terme pour en préserver l'accessibilité, nous nous pencherons sur l'archivage des données de la biodiversité : son importance, ses moyens (conceptuels et pratiques à la fois) ainsi que la situation actuelle.

I. LA TAXONOMIE : HISTOIRE ET CONTEXTE

La taxonomie (ou taxinomie, les deux termes étant synonymes et faisant d'ailleurs l'objet d'un débat quand auquel préférer, entre partisans de l'usage et partisans de la justesse étymologique : voir (Sauvage, 1965), par exemple) est une discipline de la biologie visant à identifier et définir les espèces, au moyen tout d'abord de la comparaison morphologique, puis, plus récemment, des analyses permises par les technologies numériques comme le séquençage ADN. Nous aimerions ici nous pencher sur les fondamentaux et l'historique de la taxonomie, en distinguant trois grandes périodes. La première s'étend du XVIIIe siècle jusqu'aux années 1970 – 1980, et voit la naissance et le développement de la taxonomie traditionnelle. Une seconde période, des années 1980 aux années 2000 environ, peut être considérée comme une « crise » de la taxonomie qui se retrouve confrontée à de nouvelles technologies, qui impactent sa façon de travailler, et à des enjeux de plus en plus pressants. Enfin, depuis les années 2000 – 2010, la taxonomie surmonte les obstacles pour reprendre sa place au sein de la biologie.

A) La taxonomie traditionnelle

Pour la plus grande partie de son histoire, la taxonomie s'est fondée uniquement sur l'observation et la morphologie des êtres vivants pour prouver les hypothèses scientifiques qu'elle avant. Après un bref historique, nous détaillerons la méthode morphologique, puis nous soulignerons l'importance de la discipline pour les autres champs de la biologie.

1) Les débuts de la taxonomie

La taxonomie vise à définir les espèces par l'établissement de différences entre elles. Ce n'est pas tout à fait comme la systématique, qui cherche plutôt à étudier les règles de classement des espèces sur l'arbre du vivant ; les frontières entre les deux disciplines sont toutefois poreuses, parce que l'établissement d'une espèce, ou, plus généralement, d'un taxon, nécessite également de réfléchir à sa place dans l'arbre du vivant, donc de lui attribuer une classification et de la justifier en étudiant les relations génétiques entre les différents taxons.

Les premières classifications du vivant remontent à l'Antiquité, avec notamment celles que proposent Aristote pour les animaux et Théophraste pour les plantes. Pline l'Ancien, quand à lui, fait référence pour la description des espèces. Au cours des siècles qui suivent, la classification et la description des espèces reprennent les catégories proposées par ces trois érudits antiques, puis, au Moyen Âge, on se réfère au plan de Dieu : l'histoire naturelle sert alors à décrire sa toute puissance telle qu'elle est perçue par la société médiévale (Draelants, 1996). Il faut attendre l'époque moderne pour voir apparaître à nouveau un intérêt pour la description et la classification des espèces, en parallèle avec l'essor des jardins botaniques reliés aux universités, comme ceux de Pise et Padoue (1544 et 1545), et en France ceux de Montpellier en 1593 et de Paris en 1624 (Barabé et al., 2012). La biologie, au siècle suivant, acquiert un statut de science grâce aux contributions, entre autres, de Buffon, qui publie son *Histoire Naturelle* entre 1749

et 1804, mais aussi de la dynastie des Jussieu (botanistes et médecins). La taxonomie, plus spécifiquement, va profiter surtout des travaux de Carl von Linné, qui pose les bases de la classification des espèces telle qu'on la pratique encore de nos jours : on lui attribue non seulement la hiérarchie entre les différents taxons (genre, famille, ordre, classe, règne) mais surtout le nom binominal en latin qui désigne l'espèce au sein d'un genre défini précédemment. Cette nomenclature scientifique est toujours utilisée de nos jours, avec toutefois quelques amendements.

Linné, cependant, et comme beaucoup des savants du XVIIIe siècle, considère que les espèces ne changent pas au fil du temps : elles auraient été créées telles quelles par Dieu et n'auraient jamais évolué. Les théories transformistes (portées notamment par Jean-Baptite de Lamarck) puis évolutionnistes (développées par Charles Darwin suite à son expédition sur le *Beagle* entre 1831 et 1836) vont remettre en cause cette conception fixiste des espèces et bouleverser la classification des taxons. *On the Origins of Species*, l'ouvrage phare de Darwin paru en 1859, propose ainsi une classification des espèces selon leur généalogie, et non plus seulement selon leur ressemblance morphologique. La taxonomie et la systématique se font alors dans la perspective de retrouver les liens entre les taxons, et éventuellement de remonter jusqu'à l'ancêtre commun initial. La sélection par le biais de la reproduction naturelle uniquement a depuis été remise en question, mais elle a fortement influencé la société de la fin du XIXe siècle puis du XXe, au-delà des sciences naturelles. L'avantage de la sélection naturelle telle que présentée par Darwin est de pouvoir expliquer avec élégance la morphologie des êtres vivants : ceux-ci se seraient adaptés à leur environnement au gré des mutations sélectionnées par la reproduction. C'est toutefois négliger le rôle du hasard dans l'apparition des mutations, d'autant que celles-ci sont parfois neutres ou peu bénéfiques.

Notons également que la taxonomie, au XIXe comme au XXe siècle, a fait son miel de la découverte de nouveaux territoires par les européens, et, plus généralement du contexte politique d'expansion européenne, comme l'écrit Browne en 1997 (Browne, 1997) :

> « Une fois que les collections arrivaient en Europe et qu'elles étaient placées dans un muséum à Paris, Londres ou Berlin, elles constituaient une démonstration visuelle de savoir et de pouvoir, subtile combinaison de pouvoir national, de pouvoir géographique et de pouvoir scientifique. »

Si voyager au XIXe siècle n'est pas simple, les grandes expéditions scientifiques ont été facilitées par la possession (ou l'acquisition de ceux-ci) de territoires, leur progressive mise en carte, et donc leur appropriation effective par les puissances occidentales. La posture du savant, qui nomme les choses, est aussi celle de celui qui dépossède les populations locales du savoir et de la connaissance des objets nommés. On peut ainsi considérer que la taxonomie a participé, à son échelle, aux processus de colonisation et d'appropriation des territoires et des ressources. On peut également étendre, à notre avis, cela aux exploitations industrielles (dont, plus récemment, pharmaceutiques), qui ont été faites des ressources.

Pour conclure sur ses origines, on peut dire que la taxonomie est un moyen de « fixer » provisoirement les espèces (animales et végétales) pour les classifier : elle permet ainsi le passage de la raison graphique à celui de la raison classificatrice (selon la définition de ces deux concepts par Pascal Robert (Robert,

2010)) dans l'appréhension de la biodiversité, dans la mesure où le passage de la description à la classification consacre l'espèce dans un lieu scientifique.

2) La méthode morphologique

L'image traditionnelle du taxonomiste est celle de l'homme de terrain, qui capture des animaux ou les observe aux jumelles, cueille des plantes, le tout pour les étudier et les « typifier » : de ces observations sont déduites les distinctions entre les espèces. Peut-être du fait de ce travail de terrain, qui prend en effet une grande partie du temps du taxonomiste, la taxonomie est perçue comme une discipline très individualiste, où chacun préférerait la compagnie de l'espèce qu'il étudie à celle de ses collègues, pour exagérer la caricature. Mais peut-être que la meilleure façon de représenter un taxonomiste est celle que mettent en avant dans l'introduction au dossier consacré à la taxonomie en 2004 par le journal *Philosophical Transaction of Society B* (Godfray & Knapp, 2004) :

> *« You need only look at how taxonomists are sometimes portrayed by their colleagues in other fields: as scientists who do a valuable job yet have the irritating habit of changing names for no apparent purpose. »*[1]

De fait, le travail du taxonomiste s'effectue souvent seul, chacun d'entre eux se spécialisant sur un genre en particulier, sur lequel il travaille pendant toute sa carrière. Cela s'explique par la grande diversité du vivant : il faut un œil d'expert pour déterminer deux espèces proches, surtout dans le cas de la méthode morphologique telle qu'elle a été pratiquée pendant deux siècles et demi. Il s'agit en effet de comparer les caractéristiques des individus pour en tirer des différences et des ressemblances, sur les postulats suivants : les individus partageant le plus de caractères communs font partie de la même espèce, les espèces partageant le plus de caractères communs font partie du même genre, et ainsi de suite en remontant dans la hiérarchie de la taxonomie.

La comparaison se fait à plusieurs échelles : on peut comparer ainsi la taille des taxons et des différentes parties qui les composent, leur anatomie, leurs couleurs et leurs motifs, leurs comportements, etc. L'ensemble des résultats met ensuite en exergue les différences entre les espèces et permet de les replacer sur l'arbre du vivant, que celui-ci soit phylogénétique ou cladistique.

[1] « Il n'y a qu'a voir la façon dont les taxonomistes sont parfois décrits par leurs collègues d'autres disciplines : des scientifiques qui font un travail important, mais qui ont la désagréable habitude de changer des noms sans raison apparente »

Allium filidentiforme Vved. (Vvedensky 1952, p. 32)

Type: Kyrgyzstan, Northern part of Suzak Mts, 4 July 1945, Kalinina and Moreva 71 (holotype: LE).

Perennial, bulbiferous herb without rhizomes. Bulbs ovoid, up to 2.5 cm in diameter. Outer tunics brown, reticulate; bulb neck not protruding above ground; inner tunics papery with prominent veins. Bulblets few, solitary. Stems 50–70 cm tall, 5 mm in diameter, terete, glabrous, covered with leaf sheaths for 1/3 of their length. Leaves 3–4, fistulose, spaced, blades 20–40 cm long, 3 mm in diameter, gradually tapering to apex, glabrous, green. Inflorescence 5 cm in diameter, globose or nearly so, rather lax, multiflorous. Spathe ovate, with a short beak, white-membranous. Pedicels nearly equal, many times as long as the perianth, pale green, glabrous, bracteolate at base; bracts numerous, linear or filiform, white-membranous. Perianth ovoid, indistinctly opening in flower. Tepals obtuse, apically rounded or slightly emarginated, greenish-white with a very narrow darker green median vein; outer ones broadly lanceolate, 4.0–4.5 mm long, 1.5 mm wide; inner ones linear-lanceolate, 4 mm long, 1 mm wide. Filaments white, 1.5 times as long as the perianth, with finely ciliate margin; outer ones basally triangular, apically filiform; inner ones basally ovate, broader than the outer filaments, narrowly attenuated in the upper third part to form an anther-bearing cusp, 1-toothed on each side with teeth about as long as the central cusp. Anthers pink. Pistil linear, up to 2 mm long, enclosed in the perianth. Capsule 4 mm long, 5 mm in diameter; valves rugose when dry, glabrous. Seeds compressed, black.

Phenology : Flowering in June and July, fruiting in July.

Ecology : The species occurs on sandstone deposits of various kinds, at altitudes of 900–1500 m a.s.l

Distribution : Figure 3. Kyrgyzstan, Uzbekistan (foothills and low mountains surrounding the Fergana Depression) (Vvedensky 1971, Khassanov 2017) »

Illustration 1 : Description d'Allium filidentiforme Vved. (Vvedensky 1952, p. 32) par Sennikov et Lazkov (2023). Pour la traduction, voir Annexe 1.

Cet arbre permet de retracer les liens entre les taxons, et cette technologie intellectuelle (au sens de Pascal Robert, voir plus haut) explicite les relations pour les rendre compréhensibles. La taxonomie est en fait une discipline qui se fonde sur des hypothèses de classement : on suppose, étant donné l'état des connaissances et des diverses observations qui ont pu être réalisées, que telle espèce, représentée par tel spécimen type, appartient à tel genre, et que celui appartient à tel famille, qui elle-même appartient à tel ordre, etc. Le travail de révision des hypothèses taxonomiques est alors conséquent : le travail à partir de spécimens conservés par les collections naturalistes pour préciser les anciennes descriptions est conséquent. C'est ainsi ce qu'ont fait, par exemple, Sennikov et Lazkov dans un article récent (Sennikov & Lazkov, 2023), et dont nous reproduisons ci-dessous un extrait :

On peut voir ici que la description morphologique de l'*Allium filidentiforme* Vved prend le plus de place : elle se veut précise et minutieuse pour faciliter l'identification de la plante. Les auteurs ont également inclus des photographies pour mieux visualiser les fleurs, ainsi que des cartes pour montrer la répartition géographique des plantes. À l'article est attaché un jeu de données au formalisme

DarwinCore déposé sur Dryad, dans lequel on retrouve un tableur (dans un format qui toutefois n'est pas ouvert) reprenant l'ensemble des spécimens et des observations effectuées pour réviser *Allium filidens* l.s. L'analyse morphologique de différents spécimens, ainsi que quelques observations dans la nature ont permis ici de distinguer *Allium filidens* d'*Allium filidentiforme* et de proposer pour ce dernier une description précise ainsi que des éléments de localisation.

Il arrive toutefois que la morphologie soit insuffisante pour différencier certaines espèces, et l'arrivée du séquençage ADN a permis de clarifier certaines hypothèses, mais nous développerons ce sujet plus loin.

3) Une discipline au cœur des sciences naturelles : nommer les choses

La taxonomie permet de nommer les unités qui servent de base à la recherche en biologie ; c'est également à elle que revient la tâche de classer ces unités dans l'arbre du vivant. Pour cela, elle a développé un vocabulaire qui lui est spécifique, notamment le concept de taxon : celui-ci désigne n'importe quel niveau de la classification, et ce qui est contenu à l'intérieur. Le terme a d'abord été utilisé comme fausse étymologie pour « taxonomie » par Adolf Meyer-Abich en 1926, puis par Hermann Johannes Lam en 1948 pour désigner un ensemble taxonomique, et c'est ce sens qu'il a encore aujourd'hui. Le taxon est utilisé par l'ensemble des disciplines de la biologie comme raccourci, au lieu de répéter le nom latin, toujours plus long que deux syllabes.

Un taxon désigne ainsi un ensemble de descriptions taxonomiques, qui occupent des rangs différents au sein de la taxonomie. Ces divisions et subdivisions successives permettent de classer les espèces selon leur appartenances à certains taxons : cette hiérarchie peut toutefois être affinée selon les besoins classificatoires des disciplines. Les principales catégories, par ordre décroissant, sont le règne (animal, végétal ou bactériologique), le phylum, la classe, l'ordre, la famille, le genre et enfin l'espèce

Le nom latin, lui, fait l'objet de nombreuses règles selon le règne auquel le taxon appartient. Ainsi, selon les articles 4 et 5 du *Code International de Nomenclature Zoologique (International Commission on Zoological Nomenclature et al., 1999)*, les noms des taxons au rang supérieur à celui des espèces consistent en un seul mot, en latin, qui commence par une majuscule, tandis que le nom d'une espèce est constitué de deux mots en latin, dont le premier est le nom générique (affecté d'une majuscule) et le second l'épithète (tout en minuscule). Dans le *Code International de Nomenclature pour les algues, les champignons et les plantes* (Turland et al., 2018), il est précisé que les noms de taxons supérieurs à la famille sont toujours au pluriel et commencent par une majuscule (art. 16.1), le nom d'un genre est au singulier et commence avec une majuscule (art. 20.1) et enfin, pour le nom d'une espèce (art 23.1) :

> « Le nom d'une espèce est une combinaison binaire formée du nom du genre suivi par une unique épithète spécifique qui prend la forme d'un adjectif, d'un nom au génitif, ou d'un mot en apposition. »

Les Codes donnent également des définitions sur les différents types de noms : ainsi, on distingue un synonyme (un nom donné à deux taxons différents : il peut arriver que deux noms classés dans des taxons différents soient formés de la même manière) des homonymes (deux noms donnés à un seul taxon, dans le cas où la description la plus récente n'avait pas connaissance de la précédente et a attribué un nom à quelque chose de déjà décrit). Un basionyme est le premier nom scientifique donné à une espèce, quand bien même celle-ci a changé de genre au cours de son histoire. Le choix du nom d'un taxon est laissé à la discrétion des auteurs de la description, même si quelques règles et

bonnes pratiques sont recommandées : ainsi, pour les plantes, on privilégie les mots que l'on peut facilement latiniser, de préférence pas très longs. Les scientifiques ne sont pas toutefois complètement imperméables à la culture populaire : il arrive que certaines espèces soient nommées d'après des personnalités ou avec humour (Jozwiak et al., 2015), comme par exemple la tortue *Apseudes atuini* (Bamber 2005) – d'après la tortue A'Tuin qui porte le Disque-Monde dans les romans de Terry Pratchett, auteur de fantasy britannique – ou le ptérosaure *Coloborhynchus spielbergi* (Veldmeijer 2003) pour Steven Spielberg.

Les taxonomistes nomment les espèces selon les principes édictés par les Codes de nomenclature, et classent les taxons selon le nombre de subdivisions qu'ils possèdent : ainsi, une espèce est globalement perçue comme l'unité de base, la plus exclusive, tandis que le genre inclut plusieurs espèces, la famille plusieurs genres, l'ordre plusieurs familles, la classe plusieurs ordres, le phylum plusieurs classes, et le règne plusieurs phylums. Mais cette hiérarchie des taxons a été remise en cause dans le sens où les arbres du vivant ne représentent pas systématiquement les liens généalogiques entre les espèces : il existe ainsi plusieurs types d'arbres du vivant, selon que l'on veut représenter les hiérarchies entre les taxons ou les liens évolutifs entre les espèces.

Mais puisque la taxonomie nomme, elle impose le vocabulaire aux autres disciplines de la biologie : ce travail est toutefois compris par les autres disciplines biologiques comme la définition stable de concepts ontologiques qui se rapportent à une espèce, tandis que la taxonomie émet en fait des hypothèses sur ces concepts, qui peuvent donc varier au fil du temps et des révisions taxonomiques. Ces variations ne sont pas bien perçues par les autres disciplines, qui doivent jongler avec des noms et des classifications différentes pour leurs objets d'études. La taxonomie, perçue comme une discipline de « services » par le reste de la biologie, doit alors faire face à une contradiction entre la façon dont elle se représente et la façon dont les autres la conçoivent (Barberousse & Samadi, 2013).

Cette différence entre les perceptions va s'exacerber avec l'arrivée de nouvelles technologies, qui vont changer les pratiques des biologistes. Les taxonomistes vont se retrouver face à une modernité technique qui ne va d'abord pas les intéresser, avant de se retrouver sous le feu de critiques acerbes et connaître une période de crise, entre les années 1980 et 2000.

B) L'APPARITION DU NUMÉRIQUE

1) Les sciences du vivant et les technologies numériques

Au tournant des années 1980-90, la démocratisation des technologies numériques et des outils associés a fait rentrer les sciences naturelles dans ce que nous pourrions appeler une « ère des données ». Celles-ci sont en effet de plus en plus nombreuses car issues des résultats des instruments qui servent aux scientifiques dans leurs travaux : on peut ainsi penser aux pièges photographiques, qui, en devenant de plus en plus accessibles, ont permis d'améliorer la précision, quantitativement et qualitativement, des comptages et des surveillances de

territoires sauvages. Toujours dans ces deux objectifs, le développement de balises GPS placés directement sur les animaux a permis de mieux étudier leur comportements à travers leurs déplacements quotidiens et à grande échelle, notamment pour les migrations. Mais surtout, le développement du séquençage ADN, qui permet d'étudier le vivant à une échelle encore inédite pour la fin du XXe siècle, a été le facteur de nombreuses découvertes dans le domaine de la biodiversité, en plus de faire avancer certaines théories comme celle de l'évolution.

Ces données numériques forment aujourd'hui une grande partie des connaissances que nous possédons sur la faune, et dans une moindre mesure, la flore : les plantes ont également profité du passage à l'échelle moléculaire dans la recherche pharmaceutique, qui se retrouve parfois à reproduire synthétiquement des configurations déjà présentes dans les plantes.

À l'arrivée d'Internet et la création des bases de données en ligne, est apparu le besoin de partager entre chercheurs diverses informations concernant les spécimens, et leurs observations. Parmi les différents standards qui ont été développé à l'époque pour faciliter la communication de ces informations, le Darwin Core et l'ABCD ont réussi à tirer leur épingle du jeu et sont toujours utilisés aujourd'hui.

Le Darwin Core est certainement le standard le plus répandu de nos jours, du fait de sa simplicité et sa flexibilité : reprenant les principes du Dublin Core mais appliqué aux observations et aux spécimens, il est utilisé dans la plupart des bases de données de la biodiversité. Son développement a commencé à en 1999 avant d'être reconnu comme standard en 2009 par le TDWG (Wieczorek et al., 2012). Aujourd'hui, il est notamment utilisé par le GBIF pour la description des jeux de données qui sont mis en ligne sur cette plateforme.

L'ABCD (*Access to Biological Collection Data*) est un schéma XML qui se veut structuré et structurant, et donc un peu plus contraignant que le Darwin Core. Conçu pour décrire les spécimens conservés dans les collections naturalistes, il a été développé au début des années 2000 et a été reconnu comme standard par le TDWG en 2005[2]. Aujourd'hui, il est utilisé principalement sur BioCASe et le GBIF.

Le TDWG (*Taxonomy Database Working Group*)[3] est certainement l'organisation qui a le plus fait pour le développement des standards de transmission des données de la biodiversité : elle a en effet chapeauté les deux standards mentionnés plus haut, en plus d'avoir participé à la création de plusieurs autres. Créé en 1986[4] pour travailler initialement sur les bases de données de la taxonomie botanique, le groupe de travail a vu son champ d'action s'élargir à l'ensemble de la taxonomie, avant de devenir, en 2006, le *Biodiversity Information Standard* (même si l'ancien acronyme est resté par commodité) : le but du groupe de travail était désormais de développer et de maintenir des standards pour le partage des données de la biodiversité sur internet.

Ces nouvelles technologies ont occupé (et occupent toujours, d'ailleurs) le devant de la scène de la recherche en biologie pendant plus d'une vingtaine d'années, et ce pour plusieurs raisons.

La première que l'on peut identifier est qu'elles permettent de donner au moins l'impression de mieux connaître la biodiversité, à défaut de permettre véritablement cette connaissance. Une plus grande masse de données permet ainsi d'avoir des approches quantitatives plus fiables lorsqu'il s'agit d'en tirer des tendances et des moyennes : les politiques de conservation s'appuient sur des données chiffrées, et pouvoir s'appuyer sur des ensembles toujours plus grands de données leur permet

[2]Access to biological collection data (Abcd) schema. (s. d.). Consulté 18 août 2023, à l'adresse https://www.tdwg.org/standards/abcd/

[3]Biodiversity information standards(Tdwg). (s. d.). Consulté 18 août 2023, à l'adresse https://www.tdwg.org/

[4]TDWG: History. (s. d.). TDWG. Consulté 18 août 2023, à l'adresse https://web.archive.org/web/20180329125359/http://www.tdwg.org/about-tdwg/history/

d'avoir une vision supposément plus représentative de la réalité en plus d'être perçu comme un progrès gigantesque. L'importance des travaux de recherche à grande échelle se fait d'autant plus sentir que la biodiversité est un vaste ensemble d'espèces, dont le nombre est évalué à presque 9 milliards et dont seulement quelques millions sont connues par la science. Dans ce contexte, on comprend l'utilité d'une vision englobant le plus de paramètres possibles ; mais il faut, à notre avis, également noter que les données sont difficilement exploitables d'emblée. Si le numérique permet une plus grande précision et une plus grande masse de données, il est également l'outil qui permet d'utiliser ces données, qui, de fait de leur code binaire initial, sont indéchiffrables pour un humain sans utiliser un ordinateur. Ensuite, il arrive que l'interprétation des données que fait l'ordinateur soit faussée par un manque de métadonnées techniques (format, encodage, etc), ce qui rend inexploitable les données elles-mêmes. Ainsi, accompagner les fichiers obtenus par les outils numériques de métadonnées permet aux machines aussi bien qu'aux humains d'exploiter les données qu'ils représentent : les métadonnées contextualisent la production des données et en facilitent la compréhension, aussi bien pour la machine qui peut ainsi correctement les afficher sur un écran que pour un humain qui peut mieux les appréhender en connaissant les biais des instruments de capture, par exemple.

Une autre raison de l'attrait des technologies numériques et informatiques réside en leur nouveauté. La science étant souvent perçue comme vecteur de progrès, elle se doit elle aussi d'utiliser les outils les plus neufs, ces derniers étant perçus comme « meilleurs » à tous points de vue (plus précis, plus adéquats, plus performants etc.) que les précédents. En jouant sur ce caractère nouveau, les sciences sont jugées plus attractives et c'est aujourd'hui une façon de sécuriser les financements : refaire les mêmes analyses n'intéresse pas grand monde, sauf si c'est pour mobiliser une nouvelle technologie ou une nouvelle approche des résultats. Notons que la science est également perçue comme l'exploration de nouvelles pistes qui n'avaient pas été traitées auparavant, par manque de moyens ou tout simplement d'intérêt scientifique. Mais ce n'est pas parce qu'une technologie est nouvelle qu'elle est fiable et pertinente : il y a un certain temps d'adaptation *de l'outil* et *à l'outil* pour une tâche, ce dernier temps étant de préférence moins important, puisque dans l'idéal c'est l'outil qui s'adapte à la pratique et non l'inverse. Pour les sciences naturelles, le séquençage ADN a joué ce rôle de l'outil nouveau qui s'impose à tous : cette technologie a permis de mener la recherche en biologie à un niveau alors inexploré, l'échelle moléculaire. Un autre exemple que l'on peut citer est l'utilisation de l'informatique et des traitements en masses permis par l'augmentation des puissances de calcul des ordinateurs : c'est l'arrivée des « *omics* » en biologie, qui visent à l'exploitation automatisée des résultats de recherche par les ordinateurs.

Sans minimiser les apports faits par le numérique, il faut toutefois garder à l'esprit que ce sont des technologies comme les autres, avec des avantages et des inconvénients, comme celles qu'elles viennent remplacer. Nous avons ainsi l'impression d'assister à un « mirage technologique », où le numérique est perçu comme la solution à tout.

2) Les nouveaux outils de la taxonomie

La taxonomie a elle aussi vu apparaître de nouvelles technologies pour l'assister dans la description des taxons : en fait presque toutes les technologies évoquées un peu plus haut peuvent lui être utiles.

En effet, la géolocalisation des observations et des spécimens permet de préciser les lieux et de les visualiser sur une carte avec un minimum de modification ; les balises GPS portées par les animaux permettent le suivi des populations et d'améliorer les connaissances sur les espèces, en délimitant les territoires des individus, ce qui est utile à la fois pour la conservation et la taxonomie. Les pièges photographiques permettent d'obtenir un grand nombre d'images qui peuvent être mobilisées pour la description d'une nouvelle espèce. Enfin et surtout, le séquençage ADN peut servir à faire la distinction entre deux espèces à la morphologie identique ou presque (voir par exemple la discussion dans (Parra-Olea et al., 2016))

À partir des années 1970-80, la taxonomie va surtout, en interne, revoir la façon dont elle classe les êtres vivants. De nouveaux arbres sont apparus au cours de la première moitié du XXe siècle, et deux approches (cladistique et gradiste) vont se confronter au cours des années 1970. Les arbres du vivant ont alors pour but la reconstitution de la filiation entre les taxons : le gradisme (ou la phénétique) classe les espèces selon leurs degrés de ressemblance par le biais d'algorithmes comparatifs, sans se soucier de l'hypothèse évolutive, tandis que le cladisme propose plutôt des regroupements selon le plus proche ancêtre commun.

La notion de clade – et donc le cladisme – finit par s'imposer au cours des années 1980 (Funk, 2001). Cette nouvelle façon de classer les espèces s'appuie sur la réflexion que les arbres créés jusque là ne permettaient pas de visualiser la généalogie des taxons : le but de la cladistique est alors de montrer une classification évolutive en retraçant les ancêtres communs de plusieurs espèces : on est loin de la vision fixiste des espèces proposée par Linné, qui considère la classification du vivant comme une sorte de « jardin à la française » selon l'expression de Simon Tillier (Tillier, 2005). Le terme « clade » a été théorisé par Hennig dans les années 1950, pour parler de regroupements de taxons ayant une même ascendance. Cette façon de concevoir le vivant part du postulat qu'une espèce se scinde en deux et « donne naissance » à deux espèces-filles. La possession de caractéristiques communes permet de retrouver, au sein de la biodiversité, les espèces-sœurs, puis d'en déduire leur espèce-mère. Pour cette dernière, on recommence le processus, puisqu'elle est également apparentée à une espèce-sœur et donc à une espèce-mère. On procède ainsi de suite jusqu'à remonter à l'ancêtre commun au plus grand nombre de taxons.

Le problème de cette représentation, appelée cladogramme, est qu'elle postule la disparition de l'espèce-mère au profit de ses espèces filles, ce qui n'est pas toujours le cas. De plus, les espèces peuvent évoluer en une seule espèce-fille : comment représenter cela dans un cladogramme ? Ce dernier a toutefois l'avantage de donner à voir les relations généalogiques entre espèces de façon plus satisfaisante que la hiérarchie linnéenne.

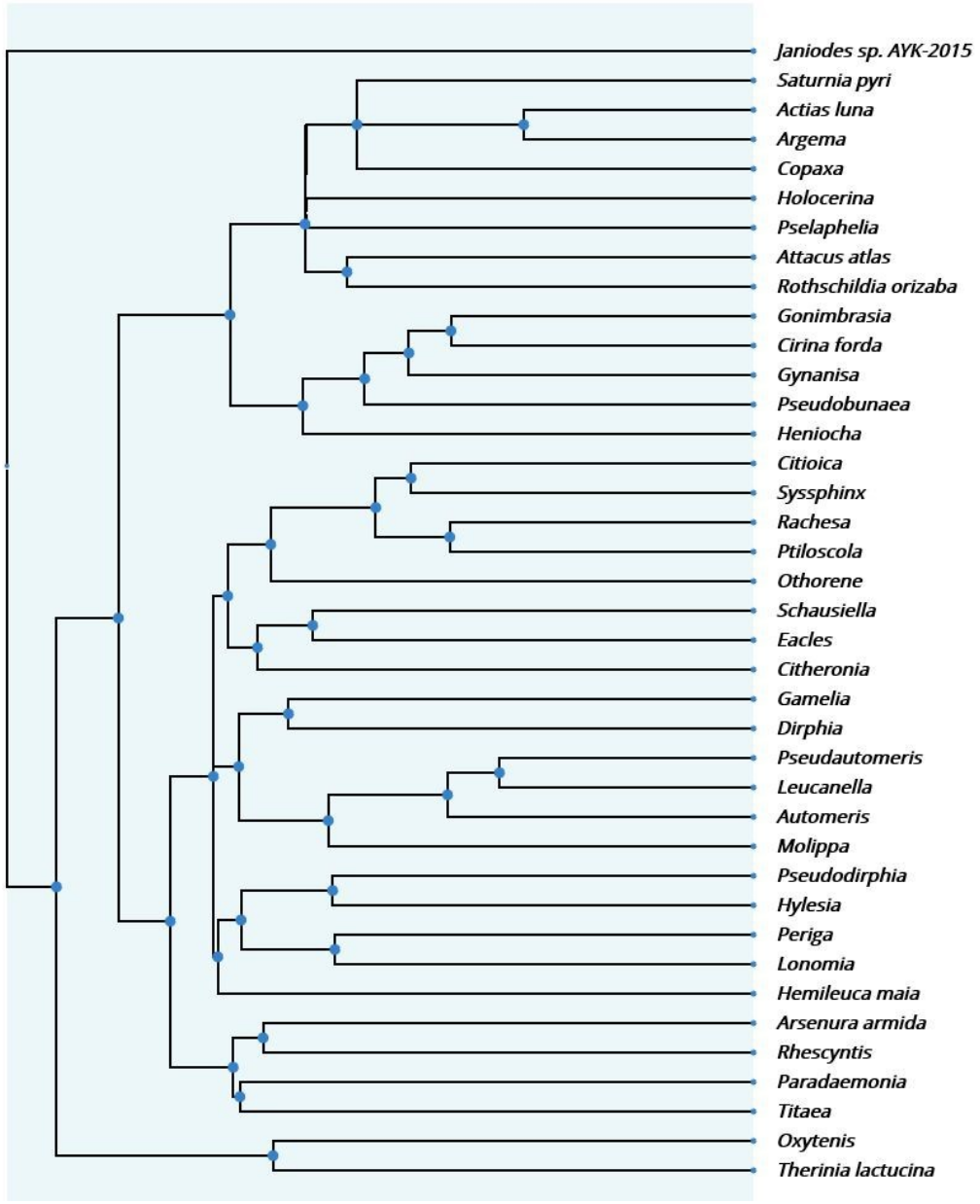

Illustration 2 : Cladogramme des Saturniidae (sous-famille des Lépidoptères, les papillons) calculée par Timetree (Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). Timetree 5 : An expanded resource for species divergence times. Molecular Biology and Evolution, 39(8), msac174. https://doi.org/10.1093/molbev/msac174
.)

La phylogénie résout ce problème en proposant une interprétation du cladogramme : elle fait apparaître le concept de « lignée taxonomique » dont les bifurcations temporelles permettent de retracer l'apparition des différentes espèces, sans préjuger de la disparition de celles-ci. Les embranchements peuvent eux aussi se ramifier et devenir à leur tour des lignées taxonomiques. Cet arbre se fonde sur la ressemblance génétique des taxons : leur proximité et leur appartenance à une même lignée est issue de la comparaison entre les séquençages ADN effectués. Les espèces les plus proches sont celles qui ont le moins de « pas » entre les différentes branche de l'arbre.

Ainsi, pour classer le vivant, la systématique élabore de nouveaux outils qui sont autant de technologies intellectuelles pour représenter les embranchements et les relations généalogiques entre les espèces. Les autres arbres qui avaient été élaborés avant la cladistique cherchaient à combiner la théorie de l'évolution et la hiérarchie linnéenne, mais, en s'essayant à cette double classification, ils étaient peu pertinents pour les deux approches.

3) Le « *taxonomic impediment* »

Malgré l'utilité que les nouvelles technologies informatiques et numériques pouvaient représenter pour eux, les taxonomistes ne s'en sont pas emparés de suite, laissant le champ libre aux autres disciplines pour le faire. Cela a mené, au fil des années, à une situation dégradée pour la taxonomie, qui a vu sa place se réduire au sein de la biologie.

L'expression de « *taxonomic impediment* », que l'on peut traduire par « handicap taxonomique », est apparue aux alentours des années 1990 pour dénoncer la situation d'alors, en s'appuyant sur un paradoxe : la taxonomie est une discipline essentielle aux autres sciences du vivant, et pourtant elle est de loin la moins financée et la plus mal en point de la biologie. Diverses raisons ont été avancée à cela, que nous allons évoquer ici.

Selon ses détracteurs, la taxonomie telle que la conçoit la méthode morphologique serait totalement dépassée. Il suffirait en effet, selon eux, de s'appuyer uniquement sur le séquençage ADN pour rapprocher deux individus et décrire une nouvelle espèce (Tautz et al., 2002). La taxonomie, grâce à cette technologie, entrerait dans la modernité et gagnerait au change, en s'alignant sur les pratiques des autres disciplines des sciences naturelles. Cette vétusté de la taxonomie lui aurait coûté sa place : puisqu'elle n'aurait pas su s'adapter à la nouvelle situation mise en place par les outils numériques, elle aurait été mécaniquement dépassée par les nouvelles approches, telles que la biologie moléculaire, qui émergeaient durant les années 1970-80.

Une autre raison au handicap taxonomique serait la lenteur avec laquelle les taxonomistes décrivent de nouvelles espèces, alors que la disparition de celles-ci s'accélère (Miller, 2007). Jugés inefficaces, les taxonomistes auraient ainsi perdu toute crédibilité auprès des biologistes, et la taxonomie en tant que discipline aurait perdu ses financements auprès des universités, puisqu'incapable d'assurer la tâche que l'on attend d'elle. À partir de cela, une sorte de cercle vicieux se serait installé : moins de postes de taxonomistes, moins de descriptions publiées, baisse de l'importance perçue de la discipline, moins de cursus dédiés à l'université, moins de taxonomistes, moins de postes ouverts etc.

La question du financement est également importante sur deux autres volets (Agnarsson & Kuntner, 2007). Le premier rejoint le problème, plus général, du financement des universités : celles-ci investissent dans les laboratoires qu'elles hébergent selon les publications des chercheurs dans des journaux à haut facteur d'impact. Or, les descriptions de taxons, par habitude, ne sont que peu citées dans les publications scientifiques, ce qui limite mécaniquement le rang des journaux auxquels peuvent prétendre les taxonomistes pour publier leurs articles. Les taxonomistes sont alors moins financés que d'autres biologistes parce que les mesures bibliographiques en vigueur ne sont pas adaptées à leurs travaux.

L'autre volet du financement concerne le coût même de la taxonomie : elle est perçue comme moins onéreuse car les taxonomistes n'auraient pas recours à des instruments de pointe. Selon ce raisonnement, un filet à insectes suffirait pour

l'équipement d'un taxonomiste qui se préoccuperait uniquement d'entomologie. C'est bien entendu négliger l'espace et les savoir-faire requis pour conserver sur le long terme les spécimens naturalisés, même si on peut objecter que cette fonction revient aux collections naturalistes. Or ces dernières sont souvent poursuivies par des objectifs de rentabilité (en nombre de visiteurs et de chiffres d'affaires) dans des pays où les musées et autres institutions culturelles relèvent rarement de politiques publiques et plus souvent d'organismes privés : elles ont ainsi pu voir leurs sources de revenus baisser au fil du temps (Kemp, 2015).

La taxonomie serait ainsi une discipline dépassée, inutile, incapable de publier dans les bonnes revues scientifiques : voilà les principales raisons du handicap taxonomique. Les implications de ce dernier sont fortes : la taxonomie serait en plein déclin, car inadaptée à l'environnement scientifique actuel. Et pourtant, elle reste extrêmement importante pour le reste de la biologie dont elle définit avec précision les objets d'étude.

Entre les années 1970 et 2000 environ, le paysage scientifique change du tout au tout en ce qui concerne la biologie. L'arrivée de nouvelles technologies, d'abord informatiques puis numériques, et surtout leur démocratisation, change les pratiques. La taxonomie a alors eu du mal à s'emparer des possibilités qui sont en train de s'ouvrir : peut-on pour autant supposer qu'elle s'est reposée sur ses acquis, sans évoluer ? C'est, selon nous, exagérer : de nouvelles façons de classifier le vivant sont apparues, comme la cladistique et la phylogénie. Mais c'est le passage à l'échelle moléculaire de la recherche et surtout son application à grande échelle qui semble montrer les limites de la taxonomie, perçue comme une discipline « de services » pour l'ensemble de la biologie.

C) Vers une taxonomie intégrative

Face aux critiques dont elle est l'objet, la taxonomie se voit contrainte de changer, au moins en partie, ses méthodes de travail et ses outils : elle en profite également pour réaffirmer sa place au sein de la communauté des biologistes.

1) L'arrivée de la méthode moléculaire

Le *barcoding* a fait son apparition dans le monde de la taxonomie au début des années 2000, dans un article de quatre chercheurs qui proposent d'utiliser comme clef d'identification pour l'ensemble du vivant le gène mitochondrial COI[5] (Hebert et al., 2003). L'article a connu un fort écho, l'efficacité de la méthode ayant initialement été démontrée pour les lépidoptères puis pour les oiseaux d'Amérique du Nord (Hebert et al., 2004). L'un des objectifs annoncés du séquençage généralisé de ce gène, tel que le présentent Costa et Carvahlo en 2007 dans la perspective du *Consortium for the Barcoding of Life* est de rendre accessible à tous l'identification des plantes et des animaux, sans avoir à passer des années à apprendre à reconnaître les caractères morphologiques de chacune d'entre elles : il suffirait d'un petit appareil, peu coûteux, et d'un prélèvement sur

[5]Abréviation de « cytochrome *c* oxidase I »

l'animal ou la plante pour séquencer son ADN et y associer une identification correcte (Costa & Carvalho, 2007).

Cette vision n'est pas sans poser quelques problèmes, tout d'abord aux yeux des taxonomistes morphologistes qui voient leurs compétences et leur travail minimisés et rendus remplaçables par des machines. Tout d'abord, l'utilisation d'un seul caractère – fut-il génétique – pour identifier une espèce est perçue comme peu fiable par la communauté des taxonomistes (Moritz & Cicero, 2004). De plus, faire du séquençage ADN la seule façon reconnue scientifiquement pour reconnaître un taxon n'est pas sans poser des problèmes éthiques, en imposant une technique alors coûteuse et peu répandue, provenant des pays du Nord qui ont la particularité d'imposer les agendas de la recherche, n'est pas sans minimiser les connaissances locales et les autres moyens d'accéder à la biodiversité : Larson postule ainsi que cela changera drastiquement le rapport du grand public à la nature (Larson, 2007) si ce dernier n'a plus à observer attentivement la faune et la flore qui l'entourent mais simplement à en prélever un échantillon.

La question du *barcoding* et de la démocratisation de l'identification du vivant qu'il amène a également été étudiée par les sciences sociales, dans la mesure où elle promeut une vision qui peut être discutable du « grand public », comme l'ont démontré Ellis et al. (2010) : ce dernier ne serait pas en mesure de s'intéresser à la biodiversité ni à l'environnement qui l'entoure sans l'intermédiaire d'un outil qui lui fournirait clef en main l'identification d'un animal ou d'une plante, outil que fournirait le séquençage à grande échelle du vivant. Or, comme l'écrivent les auteurs de l'article, c'est négliger qu'il existe d'autres communautés qui ont déjà leurs moyens de connaître la nature sans avoir à passer par le séquençage ADN.

L'arrivée du barcoding a aussi été l'occasion d'injecter de l'argent dans la taxonomie : le coût du projet du *Barcoding of Life*, dont l'objectif est de fournir un séquençage ADN pour l'entièreté du vivant alors connu, a été estimé à deux milliards de dollars (Whitfield, 2003), une somme faramineuse pour une discipline qui à ce moment voit globalement sa situation se détériorer au sein de la biologie.

L'arrivée du séquençage ADN dans la taxonomie nous semble avoir cristallisé de nombreuses tensions au sein de cette discipline, creusant un fossé entre les convaincus du séquençage et ceux qui tenaient davantage à la morphologie. On peut en effet comprendre que l'arrivée de cette nouvelle technologie ait pu effrayer des morphologistes tenants d'une taxonomie plus « traditionnelle », dans la mesure où ils se sont sentis menacés par la perte de crédibilité de leur méthode. Or, comme le reconnaissent eux-mêmes les molécularistes (voir à ce sujet l'article « Les systématiciens à l'épreuve du barcoding : Une étude des pratiques d'enrôlement scientifique » (Mauz & Faugère, 2013)), la présence de spécialistes de la morphologie des taxons est essentielle dans la mesure où il faut pouvoir être certain de l'identification d'un individu pour relier un séquençage ADN à un taxon, avant de pouvoir comparer les séquences obtenues entre elles. Loin des déclarations enflammées d'Hebert en 2003, les molécularistes ont dû revoir leur approche pour obtenir des morphologistes ce dont ils avaient besoin, à savoir à la fois les spécimens préparés de façon à pouvoir permettre le séquençage ADN mais aussi les identifications taxonomiques de ces spécimens.

Cette approche complémentaire entre taxonomie moléculaire et morphologique va se développer dans les années qui suivent les articles annonçant le séquençage comme une solution aux problèmes de la taxonomie. Nommée « taxonomie intégrative », elle se veut une synthèse des deux approches pour réinscrire l'importance de la taxonomie au sein de la biologie.

2) La taxonomie intégrative : une réponse aux critiques

Les taxonomistes morphologistes vont alors réaffirmer que leur discipline n'a pas exclusivement pour but de servir les autres champs de la biologie en leur fournissant des noms scientifiques « prêts à l'emploi », détachés de toute considération scientifique. Les noms scientifiques, rappellent-ils, sont des hypothèses de recherche, révélatrices de l'état des connaissances à un instant T, et donc susceptibles d'être infirmées ou confirmées avec le développement des connaissances (Lipscomb et al., 2003). Pour ce faire, l'étude de la morphologie de différents individus permet d'en évaluer les différences, avant d'en déduire si ces différences sont suffisamment importantes pour permettre la description d'une nouvelle espèce. Dans ce cas, le séquençage ADN seul ne peut faire office de preuve suffisante, mais son utilisation en complément d'autres informations joue le rôle d'un indice supplémentaire parmi d'autres pour l'identification de nouvelles espèces.

C'est la voie choisie par la taxonomie intégrative pour opérer une synthèse entre les deux approches, morphologiques et moléculaires. Il s'agit alors de fonder la description d'une nouvelle espèce en utilisant à la fois les techniques morphologiques et moléculaires, les deux se soutenant mutuellement dans la description scientifique du taxon. C'est une façon de concilier les deux approches et d'élever, selon nous, la taxonomie au rang des sciences « modernes », fondées sur les données et l'analyse de celles-ci.

Benoît Dayrat, dans un article intitulé « *Towards Integrative Taxonomy* » (Dayrat, 2005) a lancé le mouvement en reconnaissant les forces et les faiblesses des deux approches : il propose alors sept lignes directrices pour l'établissement de nouvelles espèces, qui définissent dans les grandes lignes la façon dont la taxonomie morphologique pourra intégrer les pratiques moléculaires à sa façon de travailler et il pose en même temps les limites à l'utilisation du séquençage ADN.

1. Tout d'abord, l'établissement d'une nouvelle espèce doit se faire après une étude approfondie des taxons du groupe où cette espèce serait classifiée ;

2. L'établissement d'une nouvelle espèce ne peut se faire qu'après avoir considéré les variations inter et intraspécifiques ;

3. Un certain nombre (à définir selon les groupes taxonomiques) d'individus récoltés est nécessaire pour établir une nouvelle espèce, mais ce nombre ne peut qu'être supérieur à un ;

4. L'abréviation sp. (pour *species*) devrait être plus souvent utilisée pour décrire un ensemble de spécimens auquel on n'arrive pas à proposer une identification correcte, avant qu'ils ne soient reconnus comme une nouvelle espèce à part entière ;

5. Les hypothèses que sont les noms d'espèce doivent s'appuyer sur un grand nombre d'éléments, parmi lesquels la morphologie est essentielle ;

6. Les spécimens holotypes doivent être conservés de façon à pouvoir faire l'objet d'un séquençage ADN par la suite ;

7. Les néotypes (c'est-à-dire les spécimens utilisés en remplacement des holotypes lorsque ceux-ci ont été perdus ou détruits) doivent eux aussi pouvoir faire l'objet d'un séquençage ADN.

Cette vision de la taxonomie intégrative nous a semblé intéressante parce qu'elle permet de donner une place égale aux deux techniques, même si on pourrait arguer que la morphologie est toujours perçue comme plus importante parce

qu'elle concerne les trois premières règles : toutefois, à notre avis, recommander que les spécimens conservés puissent permettre le séquençage ADN, c'est inscrire l'importance de ce dernier au sein même des collections naturalistes, qui sont un outil de travail très important pour les taxonomistes.

La taxonomie intégrative a toutefois mis du temps à s'imposer : cinq ans plus tard, les taxonomistes eux-mêmes ne sont pas toujours d'accord entre eux sur les moyens d'arriver à une taxonomie intégrative. José Padial et ses collègues distinguent alors deux types d' « intégrations » pour la taxonomie (Padial et al., 2010) : la première, l'intégration par congruence, se fonde sur les différences entre les lignées d'évolution des espèces (et donc la phylogénétique), tandis que l'intégration par cumulation s'appuie davantage sur les différences morphologiques pour définir de nouvelles espèces. Selon eux, les deux approches ont leurs avantages et désavantages : la méthode congruente aurait plus de difficultés à séparer les espèces tandis que la méthode cumulative aurait trop de facilités à le faire. Une façon de faire serait alors de prendre les différences morphologiques comme point de départ avant d'étudier le séquençage ADN et de voir si les résultats de celui-ci permettent de différencier deux groupes de spécimens en deux espèces différentes. Les auteurs identifient enfin quelques pistes pour développer la taxonomie intégrative : du côté scientifique, on remarque notamment l'amélioration des protocoles taxonomiques ainsi qu'une meilleure utilisation des statistiques pour la méthode congruente, et l'application des résultats de la génomique à la taxonomie. Pour ce qui est de la technique, les auteurs recommandent d'améliorer la reconnaissance automatique des espèces et le développement de logiciels pour calculer les délimitations des espèces.

3) La mise en pratique de la taxonomie intégrative

Il nous semble intéressant maintenant de voir si, et comment, les taxonomistes se sont emparés de la taxonomie intégrative dans leur travail.

De fait, les séquençages ADN commencent à se faire une place dans les descriptions taxonomiques des espèces, même si celle-ci reste limitée : pour l'année 2018, soit presque quinze ans après l'appel de Dayrat pour une taxonomie intégrative, les données moléculaires étaient fournies pour un peu plus de 90 % des publications de fungi, un peu plus de 50 % des publications de vertébrés, et entre 10 à 20 % pour les plantes et les insectes, pour un total d'environ 14 000 séquences déposées dans GenBank (Gemeinholzer et al., 2020). Ces chiffres ne concernent que les articles publiés par *Zootaxa*, qui est le journal publiant le plus de nouveaux taxons (12 % au 26 août 2023)[6]. Si ces chiffres peuvent paraître décevants, notons toutefois que les muséums d'histoire naturelle sont de plus en plus mobilisés sur la question du séquençage de leurs collections : le champ des « *museomics* », à savoir l'utilisation des collections naturalistes pour récupérer de l'« ADN historique » et de l' « ADN ancien » (Raxworthy & Smith, 2021). Du fait de l'importance des collections naturalistes pour les taxonomistes, il nous semble intéressant de considérer cette piste pour la diffusion de l'utilisation des séquençages ADN par les taxonomistes.

La faible part de descriptions scientifiques s'appuyant sur le séquençage ADN peut s'expliquer, entre autres, par le manque d'outils simples d'utilisation pour les taxonomistes. Ceux-ci commencent toutefois à se développer, à l'image de l'ensemble d'outils réunis sous le nom « iTaxoTools » par Miguel Vences et ses collègues[7]. Conçus

[6]Analytics, C. (s. d.). *ION : Index to Organism Names [Document]*. Consulté 26 août 2023, à l'adresse http://www.organismnames.com/metrics.htm?page=tsj

[7]Disponible à l'adresse https://itaxotools.org/

en langage python, les différents logiciels présents dans cet ensemble d'outils sont orientés vers l'usage des données moléculaires pour la taxonomie, et on peut les regrouper dans différentes catégories. La première simplifie la manipulation de fichiers et la standardisation de ceux-ci : on y trouve *latlonconverter* et *unitconverter* (pour harmoniser respectivement les coordonnées géographiques et les unités utilisées), *fastmerge* et *fastsplit* (fusion et séparations de séquences ADN), *dnaconvert* pour passer d'un format de séquençage à un autre, *specimentablepruner* et *specimentablemerger* pour réorganiser des fichiers autour de la colonne « *specimen* », *linebreaker* pour modifier l'encodage des sauts de lignes (qui diffèrent selon le système d'exploitation), *nodenamecorrector* pour enlever les caractères spéciaux des nœuds dans les arbres au format Newick et *spartmapper* pour convertir les délimitations d'espèces au format SPART. Une seconde catégorie contient différents programmes, utilisant tous des algorithmes différents, pour la délimitation automatisée des espèces : *PTP* (Poisson), *GMYC* (*Generalized Mixed Yule Coalescent*), tr2 (modèle bayésien), *DELINEATE* (modèle fourni par l'utilisateur), ABDG (selon les différences dans les séquençages ADN), *ASAP* (*Assemble Species by Automatic Partitioning*). Parmi les autres utilitaires présents dans le paquet logiciel, on peut noter *pyr8s* (pour calculer des arbres temporels à partir des séquences saisies dans l'interface), *TaxI2* (calcul des distances inter et intraspécifiques, *MolD* (récupération de diagnostics d'identification à partir des séquençages saisis), *dnadiagnoser* pour repérer les endroits où ont lieu les différences significatives dans les séquençages ADN, et enfin *morphometricanalyser* qui sert à effectuer des statistiques sur les caractères morphologiques présents. Le code de chacun des outils est *open-source* ; l'ensemble des modules peut être téléchargé gratuitement, ou bien être utilisé sur l'interface web mise à disposition par les créateurs du logiciel.

En mettant gratuitement autant d'outils à disposition des chercheurs, il nous semble que Vences et son équipe participent à la diffusion des outils d'analyse moléculaire à destination d'un public de taxonomistes peu familiers des outils informatiques, comme le montre la présence systématique d'une interface utilisateur, évitant ainsi de recourir aux lignes de commandes qui peuvent rebuter les non-avertis. La dimension de l'interopérabilité est présente dans la mesure où différents utilitaires permettent de convertir les fichiers et les données présentées par l'utilisateur. Nous voudrions toutefois nuancer : il n'existe, à l'heure actuelle, qu'un seul outil permettant la comparaison des caractères morphologiques. Toutefois, les auteurs de l'article présentant le logiciel à la communauté scientifique (Vences et al., 2021) indiquent que c'est une des directions à suivre pour les prochaines versions du logiciel ; ils cherchent également à intégrer davantage d'autres types de données, dont ils prévoient qu'ils seront de plus en plus utilisés à l'avenir. Il s'agit par exemple de récupérer les modèles de répartition des espèces auxquelles appartiennent les spécimens étudiés selon les occurrences qui en sont répertoriées sur la plateforme du GBIF, par exemple. Cela implique toutefois de repenser le modèle qui sous-tend le spécimen, en étendant sa portée à d'autres informations que celles actuellement collectées par les taxonomistes lorsqu'ils prélèvent un individu.

II. LE DIGITAL EXTENDED SPECIMEN

La taxonomie, on l'a vu, est au cœur des sciences naturelles. Paradoxalement, elle n'est pas très bien considérée par les autres disciplines du domaine, qui pourtant lui doivent beaucoup : si l'on prend l'exemple de la macro-écologie, qui étudie la répartition et les interactions de différentes espèces sur un territoire donné, il faut déjà être en mesure de savoir différencier les espèces les unes des autres. Pour établir une espèce, la taxonomie prélève un ou plusieurs individus, qui servent ensuite de référence pour l'espèce : on les appelle « spécimens holotypes ». Ces individus de référence peuvent par la suite être l'objet d'un reclassement, si un spécialiste se rend compte que deux spécimens que l'on pensait appartenir à deux espèces différentes n'en forment qu'une.

La notion de spécimen, que celui-ci soit holotype ou ait été prélevé au cours d'une expédition naturaliste qui avait par exemple vocation à n'étudier qu'une seule espèce, était autrefois très bien matérialisée dans les collections naturalistes. Aujourd'hui, avec l'apparition et la démocratisation des outils numériques sur le terrain et en dehors de celui-ci, la notion de spécimen se « dématérialise » elle aussi, et s'empare de ces outils pour former une nouvelle notion : le *Digital Extended Specimen* (DES), en français le spécimen numérique étendu, dont il convient de proposer une définition approfondie, avant d'en voir les usages pour les bases de données de la biodiversité, puis les avantages et inconvénients pour les chercheurs.

A. Contexte et définition

1) Contexte : les bases de données de la biodiversité

Le contexte dans lequel s'inscrit le spécimen numérique étendu est celui, beaucoup plus large, des bases de données scientifiques relatives à la biodiversité. Ces dernières sont nombreuses et couvrent à peu près tout le spectre des êtres vivants, des minéraux et des « entre-deux » (microbes et bactéries, par exemple), avec très souvent des spécialisations : fossiles de plantes, ressources ADN de vertébrés, observations ornithologiques, etc.

Or, au fil du temps, a émergé l'idée de relier ces bases de données entre elles, afin de faciliter la recherche scientifique, d'éviter au mieux les doublons entre les bases généralistes et spécialisées, et de permettre une meilleure diffusion des données. Au départ au format papier, ces bases de données ont profité de la démocratisation du Web pour exposer leurs contenus au plus grand nombre (Uhen et al., 2013). Cela n'est pas allé sans poser certains problèmes, du point de vue des chercheurs, que l'on rencontre aujourd'hui encore dans le contexte de la science ouverte.

En effet, la mise à disposition des données scientifiques est permise par des outils juridiques autant que techniques pour inciter les scientifiques à déposer leurs données, et qui visent entre autres à prévenir le vol de données et de résultats scientifiques. Les bibliothèques universitaires jouent aujourd'hui un rôle majeur dans l'accompagnement des scientifiques au dépôt de leurs données : c'est le cas par exemple des principales bibliothèques universitaires de France. Dans des domaines d'études qui peuvent être très concurrentiels du fait des forts enjeux

économiques, sociaux et environnementaux qui y sont attachés, et plus généralement dans un univers scientifique où l'évaluation des scientifiques passe systématiquement par l'étude de leurs publications et des citations obtenues, rendre ses données accessibles à tous peut être perçu comme ouvrir la porte aux réutilisations frauduleuses de la part de collègues peu scrupuleux. Notons de plus que, dans les sciences naturelles, se pose également la question de la diffusion de données sensibles, comme la localisation d'espèces protégées qui pourraient faire l'objet de braconnage : une réutilisation malveillante des données de la science doit être évitée, tout en satisfaisant aux exigences de qualité des données.

Par ailleurs, le nombre même des bases de données est un des obstacles qui peut détourner le scientifique du dépôt de ses données : devant le nombre élevé de celles-ci, où déposer ses données pour s'assurer de la meilleure diffusion possible ? Faire le choix d'une base généraliste, c'est risquer de se retrouver « noyé » dans la masse ; mais mettre en ligne dans une base spécialisée, c'est compromettre la découvrabilité des données en dehors du domaine d'expertise du producteur desdites données. Or, pour un scientifique, se retrouver face à des données légèrement hors de son domaine peut amener des réflexions et des approches fructueuses pour son propre travail.

Pour répondre à ces problèmes, les bases de données naturalistes se sont tournées vers différentes solutions. La première est d'attribuer systématiquement un auteur à un jeu de données, et de requérir l'attribution comme condition *sine qua non* de la réutilisation des données : c'est le cas du GBIF, qui distingue l'éditeur (*publisher*) des auteurs (nommés un à un). Le but ici est de donner une crédibilité et un statut scientifiques suffisants aux jeux de données pour les protéger du vol de données. Ainsi, il ne suffit pas de dire que le jeu de données est disponible sur telle ou telle plateforme, mais de pouvoir y ajouter d'autres informations essentielles comme les auteurs et les traitements effectués sur les données, par exemple. Cela passe par une description plus ou moins fine du jeu dans des fichiers descriptifs annexes, où sont notés les différents auteurs et les moyens de les contacter, la ou les institutions auxquelles ils appartiennent, les outils utilisés pour obtenir les données, etc. Un jeu de données déposé sur le GBIF indique aussi les conditions de sa réutilisation via la mention d'une licence, ici CC-BY 4.0, donc avec attribution obligatoire de l'auteur original.

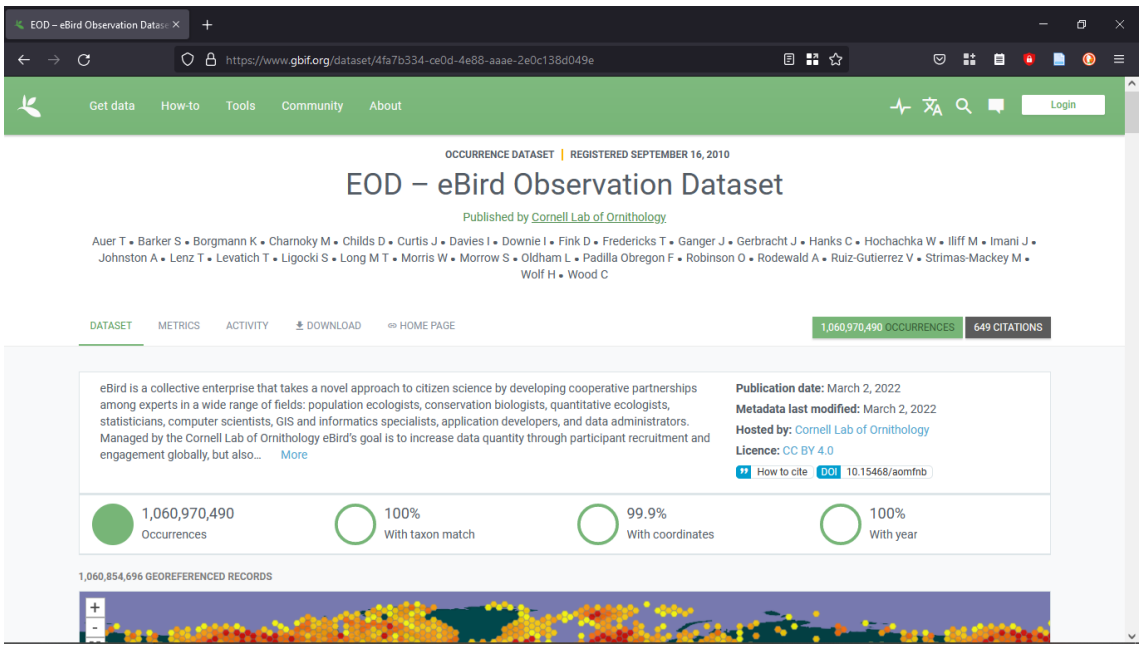

Illustration 3: Page du jeu de données reprenant les données issues de l'application eBird (consulté le 26 juillet 2023 : https://www.gbif.org/dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e). On y trouve la différence entre l'institution qui a publié les données et les personnes qui les ont élaborées, en plus de la date de publication, la date de dernière modification, l'endroit où sont hébergées les données et la licence de réutilisation des données.

Toutes ces métadonnées permettent de mieux comprendre les données présentées à la communauté scientifique et facilite alors les potentielles réutilisations : cela assure, par ricochet, une certaine forme de reconnaissance scientifique aux producteurs initiaux. Tout l'enjeu des *data papers*, qui décrivent de façon fine un jeu de données qui est disponible auprès de la communauté scientifique, est là, à notre avis : puisqu'ils cherchent à acquérir un statut scientifique similaire à celui de l'article « traditionnel », ils en présentent certains attributs comme une liste d'auteurs et un identifiant pérenne type DOI. Cela les rapproche sur la forme d'un article de recherche et donc de l'importance de ces derniers. Incidemment, signalons aussi que la description d'un jeu de données contient des éléments nécessaires pour analyser les informations *a posteriori*, comme par exemple le contexte de création des données : cela facilite leur conservation sur le long terme.

La pratique qui s'est également imposée est celle de ne mettre à disposition les données que lorsqu'elles font l'objet d'une exploitation dans un article scientifique : le lien d'accès au données et publié en fin d'article, dans un « *data availability statement* »[8]. La publication des jeux de données est alors concomitante de la publication de l'article dans lequel ils apparaissent. En plus de prévenir le vol d'idées dans un environnement parfois concurrentiel, cela ajoute de la crédibilité aux résultats publiés : n'importe quel autre scientifique peut récupérer les données, refaire les traitements proposés, ou bien tester d'autres méthodes. Dans les faits, peu de jeux sont réutilisés, sauf dans le cas de méta-analyses ; mais celles-ci sont justement permises par la publication en masse de jeux de données qualifiés et décrits correctement. Ces analyses à grande échelle sont publiées dans des articles qui citent nommément les jeux utilisés, créditant ainsi les auteurs originaux pour leur travaux.

[8]Déclaration de disponibilité des données

Enfin, relier les bases de données entre elles et les rendre interopérables s'est révélé un bon moyen pour remédier à la dispersion des données sur internet. Ainsi, il existe des portails qui moissonnent différentes bases de données pour proposer des résultats complets en évitant aux chercheurs de répéter la même requête dans plusieurs bases. C'est le cas du GBIF, par exemple, qui se présente comme un gigantesque portail permettant d'accéder à des données hébergées ailleurs. Ces processus pour relier les différentes bases de données sont parfois en cours de mise en œuvre, comme a pu nous l'indiquer Mme Pamerlon lors d'un entretien, mais l'application est facilitée par l'utilisation de standards reconnus et adoptés par la communauté des sciences naturelles.

Les principaux standards utilisés sont le Darwin Core (DwC) pour les données scientifiques elles-mêmes et l'EML pour les métadonnées qui leur sont associées. Ces deux standards, développés respectivement par le TDWG et le KNB, décrivent la très grande majorité des jeux de données. Le Darwin Core a été développé sur les mêmes principes que le Dublin Core : un ensemble de règles et d'éléments au format XML afin de décrire, relativement simplement, l'ensemble des jeux de données d'observations de spécimens de façon uniforme. Un jeu de données qui respecte le formalisme du DwC se présente sous la forme d'un fichier zippé, qui contient trois éléments : les données, au format tabulaire (csv ou parfois txt), un second fichier mappant les termes utilisés dans le premier aux termes reconnus par le DwC, et un troisième, respectant les principes établis par l'EML (parfois, l'ABCD) pour décrire les producteurs des données. L'ensemble forme l'archive Darwin Core telle qu'on peut la télécharger sur le site du GBIF.

2) Définition

Le *Digital Extended Specimen* s'inscrit ainsi dans un univers déjà normé et où l'interopérabilité est essentielle : il existe des standards déjà adoptés, reconnus et utilisés au sein de la communauté scientifique cible pour décrire les données, et les chercheurs manipulent au quotidien des données numériques décrites, pour la très grande partie, grâce à ces standards. La notion, toute récente et soutenue par l'infrastructure de recherche DISSCO (A. R. Hardisty et al., 2020), de spécimen numérique étendu intègre à la notion de spécimen telle qu'elle est comprise traditionnellement l'arrivée des pratiques numériques et le caractère étendu qu'elles amènent.

Tout d'abord, le DES reprend le concept de spécimen tel qu'il est utilisé en taxonomie et dans l'ensemble des sciences naturelles : le spécimen est l'individu d'une espèce que l'on a extrait de son environnement naturel pour l'étudier. Il sert alors de référence dans l'étude scientifique pour laquelle il a été prélevé, voire même, dans le cas d'un spécimen holotype, pour une espèce entière. Dans la forme, il peut s'agir d'un individu entier, ou bien d'une partie : les feuilles, la fleur pour une plante, par exemple. Dans le cas d'un spécimen holotype, on privilégiera toutefois l'individu entier, pour mieux appuyer la description d'une nouvelle espèce. Parmi les informations annexes qui lui sont attachées, on trouvera le nom de l'espèce à laquelle il appartient, la date et le lieu de prélèvement, ainsi que le nom du collecteur. Si le nom de l'espèce peut changer au fil des années, selon les progrès de la taxonomie, la matérialité du spécimen, pourvu qu'il soit conservé dans les conditions adéquates, peut facilement être appréhendé par les chercheurs

ainsi que le grand public : on se retrouve face à un *objet*, animal naturalisé ou feuille d'herbier, peu importe. Il est alors relativement simple de faire le lien entre le spécimen et les autres représentants de l'espèce que l'on pourrait croiser dans la nature.

Le spécimen est dit « étendu » quand d'autres informations que celles décrites un peu plus haut lui sont associées (Webster, 2018 ; Lendemer et al., 2020). Il s'agit alors de documenter les relations avec d'autres espèces : quels parasites sont présents dans la fourrure, quels insectes ont grignoté une feuille, quelle est la composition du sol, l'enregistrement du cri, une photographie de deux individus interagissant… L'objectif est ainsi d'étendre la description au-delà de l'espèce isolée, de sortir de ce cadre pour montrer toutes les interactions (ou du moins, celles qui intéressent le scientifique pour sa recherche) du spécimen dans son milieu naturel pour en retirer des informations nouvelles et pertinentes afin de mieux le connaître, notamment dans une perspective de protection et de conservation de la biodiversité. Toutefois, ces informations supplémentaires sont souvent incarnées dans des données nativement numériques.

Le spécimen numérique est, de son côté, la manifestation du spécimen physique tel qu'il est enregistré dans les bases de données des collections naturalistes ou des chercheurs eux-mêmes. Très souvent, il se présente sous la forme d'une entrée dans un tableur (ou équivalent) à laquelle est attaché un identifiant unique et un ensemble de métadonnées reprenant les informations accolées au spécimen physique. Il peut également s'agir d'une photographie d'un animal prise par une caméra automatique dans le cadre de la surveillance d'un territoire, par exemple. Le côté matériel est moins présent, même s'il faut considérer l'existence physique des serveurs et des disques durs sur lesquels sont enregistrées les données numériques : l'appréhension des données numériques reste toutefois délicate, même si joindre une image d'un spécimen numérisé peut la faciliter. Le spécimen numérique peut également être un spécimen numérisé, dans la mesure où les musées mettent en ligne leurs collections. La distinction entre spécimen numérique et spécimen numérisé se fait dans la mesure où un spécimen numérique *ne possède pas* de d'incarnation dans un échantillon physique conservé par une institution quelconque.

Faire des liens entre les différents éléments du spécimen étendu et pouvoir rediriger vers plus d'informations à leur sujet est une des possibilités ouvertes par les outils numériques. L'infrastructure technique sera détaillée un peu plus loin : pour l'instant nous préférons nous concentrer sur le concept même de spécimen numérique étendu. Celui-ci est arrivé récemment dans les sciences naturelles, pour des raisons à la fois techniques et financières : il a fallu arriver à financer des campagnes de numérisation, s'appuyer sur divers standards de métadonnées et donc les élaborer avant de les voir adoptés par la communauté scientifique… Le rôle du *DES* est de pouvoir assurer les liens et le dialogue entre différents types d'informations : entre le spécimen et sa ou ses incarnations numériques qui peuvent être de nature variée (numéro d'inventaire dans un catalogue, description au format texte, reconstitution 3D numérique, photographie haute qualité de l'objet tel qu'il est conservé dans les collections d'un musée), mais aussi entre le spécimen et les informations étendues à son sujet, comme le nom scientifique de ses parasites, l'enregistrement de sa parade nuptiale, son aire de répartition, etc. Ce processus de lien est facilité dans l'univers numérique, où il est possible, moyennant une infrastructure technique adéquate, de renvoyer vers les éléments annexes en question et de les afficher sur une même page. Comme l'écrivent Hardisty et alii (Hardisty et al., 2022):

« *a DES represents the sum of the digital information about the physical specimen, data derived from the specimen, and data associated with the specimen regardless of source* »[9].

3) Une redéfinition de la notion de spécimen ?

Toutefois, la mise en place progressive du *DES* n'est pas sans redéfinir, dans une certaine mesure, la notion même de spécimen, ainsi que ses implications pour la recherche scientifique et les collections naturalistes. Nous voudrions ici comparer certains aspects du spécimen physique et du spécimen numérique étendu.

La comparaison entre spécimens d'une même espèce est facilité par le numérique, dans la mesure où il est plus simple d'afficher les informations relatives aux exemplaires étudiés sur un même écran, tandis que la comparaison des spécimens physiques présents dans des collections éloignées est matériellement plus compliquée. Toutefois, comparer des spécimens numériques n'a de sens que si les métadonnées associées à chacun d'eux explicitent le même type d'informations, et donc suivent un standard commun. Dans le cas inverse, les informations sont inutilisables pour le chercheur qui se retrouve à comparer des données qui n'ont rien à voir entre elles. Les contraintes d'utilisation qui pèsent sur les différents types de spécimen se ressemblent sur certains points : la fragilité des objets fait écho à l'accessibilité sur le long terme des données numériques, la description d'un spécimen peut être précisée au fil du temps, etc.

Le nom scientifique est également à prendre en compte : dans l'idéal, il est fixe, c'est-à-dire qu'on ne change pas le lien entre un spécimen et le nom scientifique qu'on lui a attribué. Toutefois, les erreurs sont toujours possibles, et la science corrige parfois ses propres erreurs : une espèce peut changer de place dans l'arbre du vivant au gré des découvertes qui sont faites à son sujet. Une question qui se pose alors est celle du classement physique de l'objet spécimen, problématique à laquelle les archivistes et bibliothécaires sont régulièrement confrontés : le rangement physique doit-il correspondre au classement intellectuel ? Il n'y a pas de réponse toute faite et celle-ci doit être réfléchie au cas par cas, en fonction des capacités physiques du lieu. Cette problématique est toutefois moins présente dans l'environnement numérique, où le déplacement des ressources peut se faire plus facilement. La question se pose cependant dans d'autres termes : ce qui apparaît comme la simple mise à jour d'une notice peut s'avérer un véritable défi technique, d'autant plus si l'on souhaite conserver l'historique des informations associées à un spécimen. Il faut également s'assurer que les informations restent bien reliées les unes aux autres.

Le passage au numérique est également à envisager sous l'angle des pratiques et habitudes de recherche. En effet, la prise en main d'un spécimen physique est relativement simple, moyennant les conditions d'usage lorsque l'on manipule des objets anciens. Mais le chercheur est en terrain familier : il a fréquenté les collections naturalistes lors de ses études, sait s'y repérer, parfois même a corrigé des erreurs d'identification… Sans tout bouleverser, le passage au numérique change quelque peu la donne. La notion même de manipulation ne fait pas tout à fait sens au niveau physique, quand bien même les serveurs ont une réalité

[9] « Le spécimen numérique étendu représente la somme des informations numériques au sujet d'un spécimen physique, des données dérivées du spécimen, et des données associées à ce spécimen, peu importe leur source »

matérielle qu'il est difficile d'ignorer de part leur consommation d'électricité et le coût afférent. Toutefois cette matérialité du spécimen numérique étendu est distanciée, éloignée par les outils numériques qui la conditionnent. Ce n'est plus l'animal naturalisé ou la feuille d'herbier que le chercheur voit, mais une représentation numérique de ceci, de plus ou moins bonne qualité, avec des informations adjacentes dont la qualité dépend des moyens financiers qui ont été mis dans la numérisation. On peut comprendre que la photographie, même si elle se veut au plus près de la réalité, ne remplace pas l'objet et nécessite d'autres connaissances techniques de la part des chercheurs pour se rapprocher au plus de l'individu dans son milieu naturel.

Des reproches similaires peuvent toutefois être adressés au spécimen physique, dans la mesure où, pour les animaux naturalisés les plus anciens, les taxidermistes n'avaient pas forcément vu les animaux vivants ; pour les plantes séchées la texture et les couleurs changent avec le séchage et ne présentent plus le même aspect visuel. Il nous semble pourtant que les biologistes peuvent avoir un regard sur ces spécimens moins naïf que celui qu'ils portent au spécimen numérique, du fait de les avoir côtoyés pendant plus longtemps. Cette connaissance des faiblesses matérielles n'est peut-être pas encore atteinte pour les spécimens numériques, qui sont souvent mis en avant pour leur capacité à obtenir des données au plus près de la réalité, par leur grand nombre (Kays et al., 2020).

Notons également que la communauté scientifique exige un certain faisceau de preuves pour la découverte d'une nouvelle espèce : l'analyse ADN n'est pas suffisante, aux yeux des taxonomistes, pour accorder ce statut à un spécimen (Zamani et al., 2022). Il en est de même pour l'ensemble des données numériques reliées à un spécimen. Prises une à une, elle ne sont que des informations : mais reliées entre elles, elles forment une connaissance reliée à une espèce unique. Dans une approche « traditionnelle » de la taxonomie, c'est l'étude seule de la morphologie qui permet de déterminer une nouvelle espèce ; dans une approche « moderne », ce sont les analyses permises par les outils numériques (par exemple, le séquençage ADN) qui permettent de déterminer cela. Le spécimen numérique étendu, comme la taxonomie intégrative, se situe au croisement des deux approches. Il permet en effet de faire un lien entre les spécimens physiques et leurs informations numériques qui s'y rapportent.

Plus qu'une redéfinition, nous observons ici un élargissement de la notion de spécimen. La première étape a été celle de l'extension du spécimen à ses informations annexes, même si celles-ci étaient déjà présentes : elles sont toutefois plus précises aujourd'hui, et consignées plus systématiquement. L'ajout de l'adjectif « numérique » entérine un état de fait : à l'heure des outils numériques, qui sont aujourd'hui indispensables aux chercheurs pour mener à bien leurs travaux, cette liaison entre le spécimen et les informations qui s'y rapportent se fait dans le cadre intellectuel et technologique du numérique. Un parallèle est possible avec le concept de *distant reading* théorisé par Franco Moretti (Moretti, 2000) pour la littérature : le spécimen numérique étendu permet l'interprétation à grande échelle des données grâce aux outils numériques, sans pour autant rendre caduque la nécessité d'aller fouiller à petite échelle les bases de données.

Toutefois, il nous semble qu'un certain recul vis-à-vis du numérique, ainsi que des outils et des nouvelles pratiques qu'il amène, est nécessaire. Une certaine distance critique vis-à-vis du spécimen numérique étendu, aussi bien qu'à propos du spécimen physique, est à conserver, et une réflexion plus étendue à son sujet devrait, à notre avis, être menée par les biologistes, les taxonomistes et les philosophes des sciences.

B) Dans la pratique : utilisation par les muséums d'histoire naturelle

1) Un modèle largement partagé ?

Le spécimen numérique étendu est une notion relativement récente, et, de ce fait, on peut douter de son adoption par la communauté des biologistes. Toutefois, les différentes vagues de numérisation des collections naturalistes et l'utilisation de plus en plus importante des technologies numériques nous ont conduit à nous interroger sur l'utilisation, sinon conceptuelle, au moins en pratique, du spécimen numérique étendu par les muséums, acteurs institutionnels de la connaissance sur la biodiversité.

Nous allons ainsi passer en revue quelques collections d'histoire naturelle mises en ligne, sans pour autant prétendre à l'exhaustivité, mais pour nous faire une première idée de ce à quoi peuvent ressembler les spécimens mis à la disposition des chercheurs sur internet. La démarche se veut ici qualitative plutôt que quantitative, même si une étude plus approfondie sur les collections naturalistes à l'heure numérique serait certainement instructive. Nous avons sélectionné, sur la base de quelques recherches, cinq muséums, de petite à grande ampleur, répartis dans les pays développés, à savoir le Museum Koenig à Bonn, l'Australian Museum (Australie), le Natural History Museum (Royaume-Uni), le Smithsonian National Museum of Natural History (États-Unis d'Amérique) et le Muséum National d'Histoire Naturelle (France). Nous sommes bien conscients du biais qui est présent à la sélection de ces institutions : nous aurions aimé un panel plus représentatif, mais les deux muséums sélectionnés pour ce faire, à savoir le Muséum du Brésil et celui du Mexique, n'avaient pas, au 15 août 2023, de collections numérisées disponibles en ligne. Nous avons essayé de contrebalancer la présence de « poids lourds » du secteur en rajoutant un muséum moins connu à l'international et aux collections plus petites.

Les collections du Museum Koenig à Bonn, ouvert en 1934 et membre du *Leibnitz-Institut zur Analyse des Biodiversitätswandels*, sont riches d'un peu plus de 5 millions de spécimens[10]. Une partie des collections naturalistes a été numérisée depuis 2008 et est disponible en ligne dans un catalogue numérique qui compte un peu plus d'1,3 millions de spécimens[11]. L'interface utilisateur est simple et intuitive, et permet de rechercher un taxon selon plusieurs critères : son nom, l'institution qui le possède, sa collection, son numéro d'inventaire. Des filtres sont également disponibles, comme le pays, la collection, le (sub)phylum, la classe et l'ordre, la famille, l'espèce, le type de spécimen et la présence ou non d'un média. On peut également se promener dans l'arbre du vivant pour retrouver le taxon que l'on cherche, et savoir s'il lui est associé des séquençages ADN. Une fois le spécimen sélectionné, plusieurs informations apparaissent : tout d'abord son

[10]Chiffres calculés à partir des éléments disponibles en ligne : Museum Koenig Bonn. (s. d.). *Sammlungen | Museum Koenig Bonn*. Forschungsmuseum Koenig ; Museum Koenig Bonn. Consulté 13 août 2023, à l'adresse https://bonn.leibniz-lib.de/de/forschung/sammlungen

[11]Chiffres et dates indiqués par le muséum : Museum Koenig Bonn. (s. d.). *Digitalisierungsstrategie für die wissenschaftlichen Sammlungen am ZFMK*. Forschungsmuseum Koenig ; Museum Koenig Bonn. Et *Digital collection catalogue statistics*. (s. d.). Consulté 13 août 2023, à l'adresse https://collections.leibniz-lib.de/statistics/

numéro d'inventaire, puis son nom scientifique ainsi que sa place dans la hiérarchie linnéenne. Viennent ensuite les informations concernant le spécimen plus spécifiquement (numéro de catalogue, âge, nombre d'individus présents), les informations sur la collecte, une ou plusieurs photos du spécimen (si celles-ci existent), ainsi qu'une carte qui reprend la localisation de la collecte (dans le cas où celle-ci est géo-référencée avec précision, sinon aucune carte n'apparaît). Enfin, apparaissent la date et l'heure de la dernière actualisation de la notice que l'on consulte et les conditions de réutilisation des données présentes dans le catalogue en ligne (CC-BY 4.0), suivies de la manière de citer le catalogue.

Les collections naturalistes de l'Australian Museum de Sydney représentent plus de 22 millions de spécimens, collectés sur un peu moins de deux siècles[12]. 1,2 millions d'entre eux sont disponibles sur l'*Atlas of Living Australia* (ALA), et, d'après les filtres de recherche, au moins une image est attachée à environ 17 000 d'entre eux. Il est possible de rechercher un taxon par son nom dans l'intégralité de la base, mais comme celle-ci concentre les jeux de données de plusieurs institutions, les filtres de recherche sont particulièrement pratiques : on peut ainsi filtrer les résultats d'une requête par taxon, par date, par localisation, par type de spécimen, par la présence d'une image, et par l'origine du jeu de données. L'autre type de filtre disponible concerne la qualité des données présentées (incertitude spatiale, temporelle, etc). Lorsqu'on se rend sur une entrée du catalogue, de nombreuses données sont mises à la disposition de l'utilisateur : réparties en plusieurs catégories, elles visent à donner le plus d'informations possibles au sujet du spécimen consulté. Tout d'abord, viennent les informations au sujet du jeu de données dont le spécimen est issu : on y trouve le numéro de catalogue, l'institution et la collection qui possède le spécimen, mais également le type de celui-ci, le nom de la personne à l'avoir identifié et la date de cette identification. La licence d'utilisation des données est également spécifiée. Une seconde catégorie précise les informations liées à la collecte du spécimen : date, méthode, identifiant de l'événement. Une troisième catégorie est réservée à la place occupée par le spécimen dans la classification du vivant. Vient ensuite la localisation du spécimen lors de la collecte. La catégorie suivante rajoute des informations sur les éventuelles modifications de la notice. Trois catégories très intéressantes, de notre point de vue de professionnelle de l'information et sur le sujet du *DES*, sont ensuite présentées : les tests (réussis et échoués) sur la qualité des données, des informations sur les délimitations administratives du lieu de collecte, et pour finir, des informations sur l'environnement de collecte (températures moyennes, pluviométrie, etc), même si on ne sait pas à quelle date ont été effectués les relevés. Le cas échéant, une ou plusieurs images, ainsi qu'une carte complètent la notice. Il est aussi possible de visualiser les notices telles qu'elles ont été importées et de télécharger l'ensemble de la notice au format JSON en utilisant l'API d'ALA. Une requête dans le moteur de recherche peut également être exportée au format DwC, créant ainsi un jeu de données pour l'utilisateur.

Le Natural History Museum de Londres héberge 80 millions de spécimens, dont un peu plus de 5,5 millions ont été numérisés[13] et sont disponibles sur son « *data portal* ». Celui-ci met à disposition une API pour récupérer facilement les données, que l'on peut chercher grâce à une barre de recherche et des filtres (collection, type de spécimen, famille, genre, type de média associé et problèmes de qualité des données – plus souvent utilisé pour éliminer ce genre de données – mais aussi présence d'une image et de coordonnées géographiques). Lorsqu'on clique sur un spécimen, il est possible de visualiser la notice au format Darwin Core (qui indique quels sont les champs du

[12]Chiffres disponibles sur le site du muséum : *Natural Sciences collection areas*. (s. d.). The Australian Museum. Consulté 13 août 2023, à l'adresse https://australian.museum/learn/collections/natural-science/australian.museum/learn/collections/natural-science/

[13]Chiffres obtenus sur *Welcome—Data portal.* (s. d.). Consulté 13 août 2023, à l'adresse https://data.nhm.ac.uk/

standard qui ont été remplis) ou bien une vue un peu moins technique, mais tout autant détaillée. Ce qui apparaît tout d'abord, c'est le numéro de catalogue, suivi par la position du spécimen dans l'arbre du vivant. Vient ensuite la localisation du spécimen lors de la collecte (lieu et coordonnées géographiques) ainsi que des informations sur cette dernière (nom du collecteur, date et biome). Les informations qui suivent concernent la localisation physique du document dans les collections du NHM et le conditionnement du spécimen. Après cela sont affichés l'historique de la fiche en ligne (notamment la date de dernière modification) et les différents médias rattachés au spécimen : une carte et des images. Ces dernières sont interrogeables selon le protocole IIIF. Les trois derniers blocs d'information au sujet du spécimen consulté permettent d'obtenir un lien pérenne pour citer la notice (aussi bien la dernière version que celle que l'on est en train de consulter), mais également des liens renvoyant vers ledit spécimen dans d'autres bases de données en ligne et quelques informations additionnelles qui précisent, entre autres, les conditions de réutilisation. L'ensemble de la notice est téléchargeable aux formats DwC et JSON. Il est également possible de contacter la personne en charge de la notice pour lui signaler des erreurs. Notons par ailleurs que l'utilisateur peut lancer une recherche à partir de presque tous les champs remplis, à l'exception des éléments concernant la version numérique du spécimen (par exemple, sa date de mise en ligne ou de dernière modification).

Le catalogue en ligne du Smithsonian Museum of National History compte à l'heure actuelle un peu plus de 11 millions de spécimens numérisés, soit environ 20 % de ses collections ; environ 6 millions d'entre eux possèdent une image[14]. Lorsqu'on consulte un spécimen, les premières informations à s'afficher sont le numéro de catalogue, et le nom du catalogue dont le spécimen est issu, l'ordre, la famille et le nom scientifique du taxon, ainsi que le type de spécimen et éventuellement, la ou les citations du spécimen dans la littérature scientifique. Viennent ensuite des informations sur l'individu collecté : son sexe, son âge, la façon dont il est conservé, le lieu où il a été prélevé (à la fois sous forme textuelle et de coordonnées géographiques, si ces dernières sont présentes). Pour finir, sont affichées le nom du collecteur, la date de la collecte, la qualité des données, ainsi qu'un lien pérenne pour la notice consultée. Le cas échéant, des photographies du spécimen sont disponibles à la visualisation. Les options de recherche sont larges, et il est possible de personnaliser sa requête de façon assez étendue : chacun des champs mentionnés précédemment est interrogeable depuis l'onglet « *Specimen Inventory* ». Il est également possible de ne chercher que dans les séquençages ADN de la base de données. Le téléchargement peut s'effectuer dans deux formats différents : KML pour afficher les localisations dans Google Earth, et CSV pour obtenir les informations relatives aux spécimens. Il est possible de choisir les spécimens à exporter depuis les résultats d'une requête ; par défaut, c'est l'entièreté de celle-ci qui est exportée. Notons que les champs du tableur exporté sont simples et communs aux biologistes, mais ils ne reprennent pas textuellement les termes du DwC, et nécessitent donc un remaniement pour être réutilisables.

Terminons ce tour d'horizon des catalogues naturalistes en ligne par le Muséum National d'Histoire Naturelle de Paris. Ce dernier conserve 68 millions de spécimens dans ses collections, dont un peu plus de 10 millions ont été numérisés

[14]Chiffres obtenus sur *Si nmnh—Museum collection search*. (s. d.). Consulté 13 août 2023, à l'adresse https://collections.nmnh.si.edu/search/ et *Museum collections policies | Smithsonian National Museum of Natural History.* (s. d.). Consulté 13 août 2023, à l'adresse http://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

et sont disponibles en ligne[15], dans un catalogue qui regroupe les données d'autres institutions françaises, comme celles de l'Institut de Botanique de l'Université de Montpellier ou celles du Conservatoire et Jardin Botanique de Nancy. Les spécimens du MNHN disponibles en ligne affichent tout d'abord les informations concernant l'objet : le numéro d'inventaire, le type de conditionnement, les éventuels anciens numéros d'inventaire, l'indication portée sur l'étiquette, la détermination initiale, le sexe et l'âge de l'individu naturalisé. Une seconde catégorie, intitulée « Taxonomie », replace le taxon dans la classification linnéenne, en plus de préciser le nom commun de l'espèce, parfois dans plusieurs langues. Une troisième catégorie regroupe les informations liées à la collecte du spécimen : localisation et date de celle-ci, ainsi que le nom du collecteur. Une quatrième catégorie précise l'historique des noms scientifiques associés au spécimen consulté. Enfin, des liens vers l'espèce dans d'autres bases de données en ligne sont disponibles, et notamment vers le GBIF, même s'il arrive que les liens soient cassés, et que par conséquent il faille relancer la recherche à la main. Sur le même sujet, le lien pérenne fourni vers la notice ne fonctionne pas, sauf si l'on change le sous-domaine « coldb » par « science » (ce qui est plus cohérent avec l'adresse du catalogue en ligne). Ce lien est présenté dans une rubrique « Comment nous citer » qui est présente sur chacune des pages de spécimens, et qui propose une citation toute faite pour le spécimen. Si disponibles, les images sont affichées, ainsi qu'une carte (dans le cas où le spécimen a été géo-référencé précisément). Il est également possible de lancer une requête depuis certains champs de la notice. Par ailleurs, les résultats d'une requête peuvent être exportés au format xls, et il est possible de sélectionner les informations à exporter ; encore une fois, s'ils ne reprennent pas tels quels les termes du DwC, ils sont suffisamment clairs et communs aux biologistes pour être facilement retravaillés.

2) Analyse

Les collections numérisées que nous avons passé en revue présentent toutes, à leur façon, des spécimens numériques – et ceux-ci sont plus ou moins « étendus ». En effet, il s'agit bien d'objets numériques reprenant des spécimens présents dans les collections muséales. Toutefois, il faut noter que les informations associées à ces spécimens varient selon les bases de données : l'ALA, par exemple, est le seul à décrire précisément l'environnement de la collecte d'un spécimen ; le Koenig Museum est le seul à proposer une navigation dans un arbre du vivant ; le Smithsonian propose des fichiers compatibles avec Google Earth ; le MNHN propose de fouiller dans les catalogues d'autres institutions françaises ; le NHM propose le téléchargement de chacun de ses spécimens directement selon le formalisme du DwC. Notons en sus que tous les spécimens ne sont pas systématiquement accompagnés d'une image, alors que celle-ci est fortement recommandée par la communauté scientifique pour la représentation d'un spécimen dans l'environnement numérique. Ne serait-ce que de ce point de vue-là, aucune des cinq bases étudiées ne présente *que* des spécimens numériques étendus.

Ces derniers sont par ailleurs définis comme des objets numériques respectant les principes FAIR (Wilkinson et al., 2016), et nous allons étudier si les différentes bases de données étudiées dans cette sous-partie respectent ce critère. Pour rappel, l'acronyme FAIR englobe quatre principes (« Facile à trouver », « Accessible », « Interopérable », et « Réutilisable »).

[15]Chiffres obtenus sur *Qu'est-ce que le Muséum ?* (s. d.). Muséum national d'Histoire naturelle. Consulté 13 août 2023, à l'adresse https://www.mnhn.fr/fr/qu-est-ce-que-le-museum

a. Facile à trouver : les catalogues en ligne ont tous été trouvés en quelques clics depuis le site de l'institution à laquelle ils appartiennent. Au sein même des catalogues, la recherche de spécimens précis est facilité grâce à de nombreux filtres, pour peu que l'utilisateur soit un minimum averti du vocabulaire taxonomique.

b. Accessible : les spécimens sont consultables en ligne, sans verrou d'accès ni création de compte. Le téléchargement des données se fait tout simplement, sans compte non plus (à l'exception notable de l'ALA, qui exige la création d'un compte utilisateur gratuit, dans le but de pour récupérer des statistiques sur la finalité des jeux de données téléchargés)

c. Interopérable : différents critères sont à étudier ici. Le premier concerne les formats de fichiers que l'utilisateur peut obtenir : à l'exception sur MNHN (format XLS), le csv domine la partie. C'est un format ouvert, et donc décodable par l'ensemble des machines sans trop de difficultés. Un second critère se concentre sur l'adhésion à un standard commun, à savoir, dans le monde de la biologie, le Darwin Core ou bien l'ABCD : ces deux standards sont considérés comme des références par les biologistes. Or, seul le NHM propose l'export d'un seul spécimen selon le formalisme du DwC ; les autres institutions proposent des champs qui leur sont propres, même si ces derniers se rapprochent sur le principe des termes du DwC. Notons toutefois qu'il est possible d'obtenir des fichiers toujours en anglais, ce qui simplifie le travail du chercheur pour harmoniser les noms des colonnes des fichiers qu'il retravaille. Enfin, notons que certaines bases proposent une API pour récupérer leurs données : il s'agit de l'ALA et de NHM. Si celle-ci requiert davantage de connaissances techniques, elle permet toutefois l'automatisation pour récupérer de grands jeux de données.

d. Réutilisable : au-delà de la compréhension des données par un humain aussi bien que par une machine (et donc leur facilité à être réutilisées sans trop de manipulations techniques), les licences et conditions de réutilisation des données présentes dans les catalogues sont affichées dans la plupart des cas. Le Koenig Museum précise en bas de page que l'attribution des données au muséum est requise (licence CC-BY 4.0) ; le Smithsonian précise dans un document annexe que l'attribution est appréciée mais non obligatoire (ses collections tombent soit dans le domaine public, soit dans le *fair use* américain pour des réutilisations non-commerciales) ; le MNHN ne précise rien mais indique systématiquement la façon de citer un spécimen, suggérant ainsi fortement de le faire ; le NHM encourage très fortement à citer, sans pour autant le requérir pour la réutilisation de ses données ; enfin, l'ALA rappelle la licence de réutilisation sur chaque notice (le plus souvent, du CC-BY 3.0). Notons que ces deux dernières institutions proposent aussi un DOI unique à chaque personne qui télécharge une partie de leurs bases de données, ce qui facilite l'identification et la réutilisation des jeux.

Les spécimens numériques étendus proposés par les cinq catalogues étudiés respectent globalement les principes FAIR, mais on peut douter du caractère véritablement étendu des données proposées. Ainsi, par exemple, seul l'Australian Museum propose une description de l'environnement de collecte, et les relations entre différentes espèces ne sont mentionnées nulle part. Les images ne sont pas toujours présentes, et il n'y a pas non plus d'enregistrements concernant les individus. En revanche, il y a parfois des liens vers d'autres bases en ligne qui servent de référentiel, et peuvent ainsi agir comme « hub » de données pour une espèce. C'est par exemple le cas du GBIF, qui centralise différentes observations et

renvoie vers les différentes occurrences d'observation disponibles dans les jeux de données déposés sur les différentes plateformes qu'il moissonne.

On peut toutefois considérer que les muséums étudiés ici mettent en ligne des spécimens numériques étendus dans la mesure où ils proposent un référentiel des espèces présentes dans leurs collections : si (et c'est un grand enjeu des années à venir, à notre avis) les chercheurs prennent l'habitude de citer les catalogues en ligne et de faire référence aux spécimens numérisés comme ils le font pour des articles, il sera possible pour les muséums de rajouter les liens vers les articles en question pour enrichir leurs propres catalogues, notamment dans le cas des révisions taxonomiques (Groom et al., 2017).

3) Réflexions sur les résultats

Les muséums d'histoire naturelle dont nous avons examiné les catalogues en ligne ne semblent pas tout à fait s'inscrire dans une démarche de mise à disposition de spécimens numériques étendus. Plusieurs raisons nous semblent expliquer la situation.

Tout d'abord, rappelons que le concept de spécimen numérique étendu est neuf. Il n'en existe pas encore de réalisation pensée comme telle dès le début : les collections naturalistes numérisées appliquent mettent davantage en œuvre le concept de spécimen numérique, un peu moins récent. Cela tient à la façon dont elles ont été conçues : on peut analyser les collections mises en ligne comme une réplique dans l'environnement numérique des collections physiques, qui mettent au cœur de leur démarche l'individu naturalisé, et non ses liens avec d'autres espèces ou des enregistrements sonores ou visuels. Ainsi, puisque ces informations n'ont pas été collectées en même temps que les spécimens concernés (qui remontent pour certains au XIXe siècle), il n'est pas possible de les rajouter au moment de leur mise en ligne ; avec le changement climatique, indiquer les moyennes actuelles de températures ne permet pas de se représenter de façon fiable l'environnement de collecte, *a fortiori* lorsque la localisation initiale est imprécise (il arrive que l'on mentionne simplement le pays de collecte…). Le chercheur se retrouve face à un manque d'informations qui peut être crucial lorsqu'il cherche à comparer deux spécimens entre eux, qu'il s'agisse de deux taxons différents ou non. Pour les spécimens les plus anciens, proposer une version étendue ne peut pas passer par la diffusion d'enregistrements, mais par l'ajout des publications où le spécimen est cité dans la littérature scientifique. La mise en ligne des collections naturalistes permet toutefois de rendre accessible ces spécimens à davantage de chercheurs, et constitue une première étape dans la création d'un spécimen numérique étendu. Même si les liens ne sont pas encore faits par les machines, un humain peut comparer les spécimens conservés dans différentes institutions depuis son ordinateur, puis aller les observer « en vrai ». Le lien entre l'objet numérique et le spécimen physique est assuré par le numéro de catalogue, qui sert d'identifiant aux deux unités intellectuelles : il n'est donc pas unique, et il nous semble qu'on touche à un problème très intéressant avec cet aspect. L'identifiant ambigu est la façon la plus simple de relier les deux objets, mais elle opère une fusion entre les deux dans la façon dont les machines comprennent les choses. De la même façon qu'un nom d'espèce unit deux spécimens qui possèdent des historiques de conservation différents, l'identifiant unique entre notice numérique et spécimen physique rend la différenciation entre les deux objets difficiles. On peut se demander à quel point, avec le développement des bases de données naturalistes, les frontières vont se brouiller entre spécimens physiques, spécimens numérisés, et spécimens nativement numériques, dans les années à venir. On peut également se poser la question du

« doublon » entre ces trois éléments : le ou lesquels vont être plébiscités par les scientifiques, et dans quelle mesure ? S'il nous semble improbable que les spécimens physiques disparaissent entièrement (notamment parce que la taxonomie refuse, pour des raisons qui sont bien étayées, de passer à un spécimen uniquement numérique), les collections naturalistes devront néanmoins prendre en compte cette évolution des pratiques de leurs publics.

On peut également se questionner sur les capacités des institutions naturalistes actuelles à remplir le « cahier des charges » des bases de données dédiées aux spécimens numériques étendus. En effet, ces dernières impliquent de grandes capacités de stockage dédiées aux différents médias (images, vidéos, sons) qui peuvent s'avérer gourmands en place, dès qu'il s'agit de fichiers de bonne qualité, condition évidente lorsqu'il s'agit de pouvoir analyser finement un individu. Le volume des données est déjà conséquent par le nombre d'entrées à créer pour chacun des spécimens possédés (rappelons d'ailleurs qu'aucun muséum, parmi ceux passés en revue, ne propose actuellement l'intégralité de ses collections en ligne) ; l'ajout de fichiers supplémentaires alourdit la base et peut ralentir les temps de réaction de celle-ci lorsqu'une requête est lancée, et ce d'autant plus qu'il est d'usage de mettre à disposition plusieurs vues du spécimen s'il s'agit d'un animal naturalisé. Relier entre eux les différents éléments du spécimen numérique étendu nécessite également que chacun d'eux soit identifié précisément, de façon pérenne, et que la base soit capable de faire appel à chacun de ces éléments pour les présenter de façon conjointe sur une même page. Là encore, la question du temps de réponse est cruciale pour l'expérience utilisateur : une base de données qui possède toutes les informations importante mais qui met longtemps à afficher les résultats sera, à notre avis, moins souvent utilisée qu'une base moins précise mais plus rapide. Ce problème de latence peut bien sûr être résolu techniquement en augmentant le nombre de serveurs, mais on touche là à ce qui nous semble être le nerf de la guerre : le financement, si crucial dans ce genre de projets. Il faut ainsi prendre en compte aussi bien l'aspect matériel des bases de données (serveurs, site web, gestion des données mises en ligne aussi bien que des différents utilisateurs potentiels…) que les ressources humaines (postes d'informaticiens, de gestionnaire des collections numériques, etc) dédiées à la mise en ligne et le maintien des collections numériques sur le long terme. On peut ainsi se demander quelle part de leur budget les muséums d'histoire naturelle peuvent raisonnablement mettre à disposition de leurs plateformes internes qui hébergeraient des *DES*, sans compromettre leurs autres activités, notamment de recherche mais aussi de transmission culturelle et scientifique.

Dans ce contexte, l'arrivée d'infrastructures telles que DISSCO, dont le but explicite est de débloquer des financements pour la mise en place de bases de données dédiées aux spécimens numériques étendus, arrive à point nommé pour faciliter la diffusion du concept et permettre la création des bases de données, que ces dernières soient alimentées par les institutions ou les chercheurs eux-mêmes. Il serait toutefois intéressant, à notre avis, que les muséums prennent en compte la dimension numérique des spécimens qu'ils hébergent, surtout lorsque ceux-ci ont été collectés dans les dernières années, qu'ils ont fait l'objet d'un traitement numérique et que ces données ont été déposées en ligne. Cela peut tout à fait rentrer dans leur prérogatives, dans le sens où les collections naturalistes servent déjà de référence aux biologistes pour la morphologie des taxons et, qu'ils hébergent de plus en plus des collections de séquençage ADN ; passer au numérique permettrait de rendre disponible au plus grand nombre leurs collections

en plus de décrire plus en détail les spécimens conservés, et par conséquent la diversité du vivant. Par ailleurs, il nous semble nécessaire de poser la question de la centralisation de ce type de collections : les ressources pouvant être dispersées à plusieurs endroits sur le web, une plateforme unique permettant de rechercher à la fois les spécimens physiques, leurs incarnations numériques, et les données reliées serait certainement un plus du point de vue utilisateurs. Mais, au fait de l'écart des compétences techniques entre conservation sur le long terme de spécimens naturalisés et de données numériques, il est envisageable que cette gestion des collections puisse être menées par des personnes (voire des institutions) différentes.

Il nous semble nécessaire que, dans les années à venir, il y ait de nombreuses réflexions autour du spécimen numérique étendu et de ses implications pour les institutions d'histoire naturelle. Après avoir brièvement étudié la question sous l'angle institutionnel, nous aimerions nous pencher sur le point de vue des scientifiques qui ont vu leurs pratiques évoluer durant les dernières décennies.

C) Le spécimen numérique étendu du point de vue de la recherche

1) Faciliter la recherche scientifique

La mise en ligne des données de la biodiversité et le rassemblement des données à un même endroit permettent de simplifier la recherche d'information par les chercheurs lorsque ceux-ci ont besoin de trouver des données qui sont souvent dispersées dans plusieurs bases pour leurs travaux : le temps passé à rassembler des informations est alors moins long lorsque toutes les informations sont rassemblées au même endroit. Une étude menée récemment estime que le GBIF, en proposant un service centralisé pour la recherche des données de la biodiversité, a permis d'économiser 845 000 heures, ce qui correspondrait à 35 millions d'euros (Deloitte Economics, 2023) pour l'année 2021. Comme nous l'a rappelé lors d'un entretien Sophie Pamerlon, ingérieure données au GBIF, ces économies de temps et d'argent ont été rendues possibles grâce à la standardisation des données : rendre les jeux de données disponibles au format Darwin Core, plébiscité par la communauté des biologistes, évite d'avoir à retravailler systématiquement les données téléchargées.

La mise en ligne de spécimens numériques étendus, aussi bien pour les chercheurs que pour les collections naturalistes, permet de respecter les principes FAIR lors de la diffusion de leurs données, dans la mesure où un spécimen numérique étendu doit remplir les critères représentés par cet acronyme. Il faut toutefois noter que la mise en conformité des données avec ces principes n'est pas toujours évidente. Cela peut mener à des erreurs de la part des chercheurs, surtout lorsque ces derniers commencent leur carrière et ne sont pas au courant des bonnes pratiques comme l'expliquent de jeunes chercheurs au cours d'une conférence (Hansen et al., 2022). Les présentateurs plaident alors pour une meilleure intégration des données aux cursus de recherche et une plus grande reconnaissance des publications de jeux de données ; nous voudrions ajouter qu'ils devraient pouvoir se tourner également vers les services d'aide à la recherche, qui,

du moins en France, peuvent être compétents pour aider les chercheurs à ouvrir leurs données.

Lors d'une présentation donnée lors des journées TDWG 2021, Michael Webster présente différents avantages liés au spécimen numérique étendu (M. Webster et al., 2021) : nous en résumons ici les quatre points principaux. Tout d'abord, le spécimen numérique étendu permet d'améliorer, de façon générale, la qualité des données : relier les données dérivées ou associées au spécimen physique dont elles relèvent, c'est aussi redonner du contexte à ces données obtenues dans un second temps et les rendre plus fiables aux yeux des chercheurs, pour qui le « *voucher* », l'échantillon physique, et donc le spécimen conservé reste la meilleure preuve scientifique. Dans un second temps, Webster présente la possibilité des co-analyses, entre par exemple les parasites et les hôtes, mais également des relations entre les comportements et la morphologie de spécimens étudiés. Il mentionne ensuite qu'il est possible de relier les spécimens numériques étendus à d'autres jeux de données, avec des conséquences non négligeables, notamment pour la conservation : l'un des exemples donnés est celui de défenses d'éléphants de contrebande, saisies à la frontière, et dont l'analyse génétique a permis de retrouver le lieu où habitaient les pachydermes, permettant ainsi d'arrêter les braconniers qui avaient tué les animaux. Enfin, la quatrième application du spécimen numérique étendu présentée par Webster concerne les nouvelles applications : il mentionne ainsi l'automatisation de la reconnaissance des caractères morphologiques grâce à l'apprentissage automatique, qui pourrait se nourrir de l'ensemble de données confirmées par le savoir des taxonomistes. Il termine en notant que les différents exemples qu'il a présenté pourraient se généraliser grâce au développement du spécimen numérique étendu, en comparant la situation actuelle à celle d'un iceberg : les applications actuelles ne seraient, selon lui, que la pointe de celui-ci, et que la partie encore immergée de l'iceberg sera exploitable une fois le concept de spécimen numérique étendu généralisé.

Le spécimen numérique étendu voit ainsi ses applications en dehors de la taxonomie, et pourtant il relève, initialement, de l'expertise de cette dernière. Il nous semble que, d'une certaine manière, il représente une autre façon de faire de la taxonomie intégrative : en collectant les informations « annexes » à un spécimen et en les mettant à disposition des autres disciplines de la biologie, la taxonomie permet tout d'abord une description des taxons plus détaillée qu'auparavant, et peut également se nourrir d'autres approches et d'autres points de vue scientifiques. Il nous semble alors, et ce serait à creuser plus en avant, que dans la mesure où le spécimen numérique étendu permet de réunir un ensemble d'information le plus complet possible, il peut servir de référence pour les autres champs de la biologie.

2) Les désavantages du spécimen numérique étendu

Cependant, le spécimen numérique étendu n'a pas que des avantages. Nous aimerions lister ici quelques unes des difficultés qui peuvent surgir à la mise en pratique d'un tel concept.

Telle qu'appliquée actuellement par les institutions naturalistes dans la numérisation de leurs collections, la notion de spécimen numérique étendu, nous l'avons vu, ne correspond pas tout à fait à sa définition. Comme nous l'a indiqué

lors d'un entretien Aurélien Mirallès, taxonomiste à l'Institut de Systématique, Évolution, Biodiversité (UMR 7205 – CNRS – MNHN – EPHE-PSL – Université des Antilles), les procédures de numérisation ont été faites « en masse », et ne présentent parfois pas suffisamment d'informations visuelles pour permettre une identification définitive à partir de ces seules informations. Selon lui, l'approche pour la numérisation a été plus quantitative que qualitative : il se montre alors insatisfait de la réalisation du spécimen numérique étendu, qui selon lui aurait dû lui permettre d'identifier une espèce sans avoir à bouger de devant son ordinateur. De fait, la numérisation des spécimens est coûteuse, et dans la mesure où les collections contiennent des millions de spécimens, le budget dédié à la numérisation haute qualité, avec plusieurs images par spécimens, est conséquent. Pour mettre en ligne et de façon plutôt rapide leurs collections, la numérisation des collections patrimoniales s'est ainsi concentrée, il nous semble, sur l'automatisation des processus (Tegelberg et al., 2014).

De plus, notons qu'une numérisation des collections sans vérification et mise à jour des identifications ne peut être que partielle et surtout potentiellement incorrecte, ne facilitant pas le travail des biologistes : ces deux étapes doivent donc être intégrées aux différentes étapes de la numérisation (Nelson et al., 2012). Sur le même thème, il nous semble souhaitable que les métadonnées associées aux spécimens numériques étendus soient correctes tout d'abord, mais également cohérentes avec les standards déjà utilisés par les biologistes, afin de respecter l'interopérabilité et la réutilisation promues par les principes FAIR auxquels les spécimens numériques étendus doivent être conformes.

La centralisation des informations est à la fois un avantage et un inconvénient du point de vue de la conservation des données : si toutes les informations étendues d'un spécimen numérique sont hébergées par une même institution, il suffit que cette dernière n'ait plus les moyens matériels et humains de l'entretenir pour que la base de données constituée par les collections numérisées disparaisse. L'avantage du spécimen numérique étendu est que les différentes informations qui le composent peuvent être conservées par plusieurs institutions ou dans plusieurs entrepôts de données différents, ce qui a le mérite de répartir la charge entre plusieurs acteurs ; il faut toutefois que l'entente entre ceux-ci soit bonne pour maintenir les liens à jour et proposer un ensemble de données cohérent. L'existence de « super-catalogues » moissonnant diverses bases de données pourrait permettre la décentralisation des données disponibles, tout en gardant une interface utilisateur simple. Mais cela pose des problèmes techniques, comme nous l'a confirmé Mme Sophie Pamerlon, ingénieure données au GBIF.

De plus, et même si nous avons déjà abordé brièvement la question, il nous semble important de préciser que les financements pour ces bases de données doivent être pensés sur le long terme : les informations présentées devraient dans l'idéal être régulièrement mises à jour, ne serait-ce que pour les noms scientifiques, qui, on l'a vu, peuvent changer au fil des hypothèses taxonomiques.

Enfin, un des effets de l'utilisation généralisée des spécimens numériques étendus serait, à notre avis, de réduire la consultation physique des collections naturalistes : puisque tout serait disponible en ligne, le chercheur n'aurait plus besoin de consulter les collections physiques pour faire son travail. Si en soit ce n'est pas une mauvaise chose – pensons notamment à la consultation des spécimens par des membres d'institutions n'ayant pas suffisamment de moyens pour permettre à ses membres de voyager – cela change toutefois la façon de faire de la recherche. L'un des enjeux pour les collections muséales sera alors d'obtenir des chercheurs qu'ils citent les spécimens numériques mis en ligne, et reconnaissent l'importance de ceux-ci au sein de leur travail, afin de ne pas « couper les vivres » à des collections naturalistes qui seraient moins fréquentées. Il est toutefois plus probable, à notre avis, que les scientifiques continuent à mobiliser les deux formes du spécimen, du moins tant que le spécimen numérique reste en partie

insatisfaisant : il nous paraît difficile de reproduire les sensations tactiles, par exemple, grâce à des représentations 3D, alors que ces dernières sont déjà rares dans la mise en ligne des collections naturalistes.

Un autre élément sur lequel nous aimerions attirer l'attention est que les chargés de collections naturalistes ne sont que peu, de nos jours, formés sur la question des données, que celles-ci proviennent de numérisation ou bien soient nativement numériques. La conférence donnée par Anna K. Monfils lors des journées TDWG 2021, le montre bien : intitulée « *Workforce Capacity Development and the Digital Extended Specimen* », elle appelle à un état des lieux plus poussés sur les compétences des gestionnaires de collections naturalistes en ce qui concerne les données numériques (Monfils & Ellwood, 2021)[16]. L'une des questions posées par l'un des auditeurs par la suite est très intéressante et résume bien le carrefour auquel la biologie est actuellement parvenue : faut-il former des biologistes à la question des données ou bien former des *data-scientists* aux enjeux et pratiques particuliers de la biologie ? La question reste ouverte, et il nous semble que les archivistes et les bibliothécaires pourraient également apporter un éclairage voire une expertise sur la question, dans la mesure où ils sont de plus en plus régulièrement confrontés, eux aussi, à des documents numériques et à des problématiques de mise en ligne de ceux-ci.

3) Le spécimen numérique étendu, un moyen de redocumenter le spécimen ?

Nous souhaiterions maintenant aborder le spécimen numérique étendu du point de vue des sciences de l'information, car il nous semble que celles-ci peuvent apporter un éclairage sur la question du spécimen en tant que document.

Il nous semble ainsi qu'il faut prendre en compte les enjeux liés à la transformation des données : comme l'écrit Sabina Leonelli (Leonelli, 2010), la diffusion des données par l'intermédiaire d'une base de données se fait dans un processus de décontextualisation / recontextualisation des données. Cette façon de voir peut, à notre avis, être rapprochée de la vision d'un document telle que la conçoit Suzanne Briet dans son ouvrage *Qu'est-ce que la documentation ?* L'exemple bien connu présenté par la bibliothécaire est celui d'une antilope qui, tant qu'elle reste dans son milieu naturel, n'est pas considérée comme autre chose qu'une antilope ; mais dès qu'elle est déplacée dans un zoo, elle devient un « document », témoignage d'une réalité qui n'est pas celle où elle est exposée, et d'où l'on peut tirer des informations supplémentaires (Briet, 1951). C'est le même principe qui, selon nous, apparaît dans le chapitre de Sabina Leonelli dédié à la transformation des données pour leur mise en ligne, l'autrice faisant pour sa part le parallèle avec les collections muséales : les données telles que produites par les scientifiques doivent être sorties de leur contexte initial pour pouvoir « voyager » entre les différentes utilisations qui en seront faites. Ce processus de documentarisation des données, c'est-à-dire leur transformation en documents exploitables, ne se fait pas sans perdre des informations présentes au départ, même si la présence de métadonnées permet de pallier en partie le problème. Le spécimen numérique étendu peut être analysé comme une seconde étape dans le processus de

[16]Enregistrement vidéo disponible ici : *TDWG 2021 : SYM07 Digital Extended Specimens*. (2023). Consulté 25 août 2023, à l'adresse https://www.youtube.com/watch?v=jpY7vh7BgKM à 01:13:53

documentarisation : la première consiste en la collecte du spécimen et son intégration dans les collections d'histoire naturelle, et la mise en ligne de ces collections comme la création d'autres documents, incarnations des premiers dans l'espace numérique. Mais ces deux étapes ne se font pas sans induire des pertes d'informations : le spécimen conservé dans une collection naturaliste diffère de l'individu vivant dans son milieu naturel, ne serait-ce que parce qu'il n'est plus en vie, tout simplement. Sa conservation sur le long terme implique un changement de support : une fleur est par exemple attachée à une feuille de papier pour former une page d'herbier, un animal est conservé dans du formol ou empaillé. Les informations qui l'entourent alors sont relativement peu nombreuses : date et lieu de collecte, nom du collecteur, identification de l'espèce sont les seules choses que l'on sait à son sujet. Toutefois, la présence de ces informations joue un double rôle : d'une part, elles permettent dans une certaine mesure de recontextualiser l'individu avant qu'il ne soit considéré comme spécimen, et d'autre part, elles assurent que le spécimen puisse jouer son rôle de référence scientifique auprès des biologistes et plus particulièrement des taxonomistes.

Dans cette perspective, la mise en ligne des spécimens change à nouveau leur support : on passe d'un document « matériel », manipulable, à un document numérique, avec ce que cela peut induire d'instabilité, mais également de possibilités de copies et d'exemplaires supplémentaires. Du fait de ce changement de support, le spécimen voit la manière dont il est appréhendé changer : le scientifique n'a plus que des photographies devant lui, dans le meilleur des cas des reconstitutions 3D (mais celles-ci sont rares : nous n'en n'avons pas rencontré durant notre exploration des collections numérisées) – le chercheur perd alors des informations tactiles aussi bien que visuelles sur le spécimen. Mais, au-delà du changement de support, la création de liens vers les autres ressources qui constituent le spécimen numérique étendu peut constituer, selon nous, un processus de recontextualisation dans la mesure où le spécimen numérisé est considéré comme faisant partie d'un ensemble plus grand d'informations qu'il est possible de fournir à l'utilisateur par l'intermédiaire de l'infrastructure numérique : on regagne ainsi ce qu'on avait pu perdre à la collecte, à savoir, par exemple, l'environnement précis, les interactions avec d'autres espèces, le comportement, etc.

Il faut toutefois, et nous revenons aux travaux de Sabina Leonelli sur ce sujet (Leonelli, 2010; Leonelli & Ankeny, 2012), prendre en compte que les processus de dé-contextualisation et re-contextualisation évoqués ici changent la façon même dont on parle des spécimens : tout d'abord, la standardisation du vocabulaire peut poser problème dans la mesure où toutes les disciplines ne partagent pas les mêmes concepts – à cet égard, la flexibilité permise par le Darwin Core, format qui n'a été pensé pour aucune discipline de la biologie en particulier mais pour documenter les occurrences d'observations, a certainement permis la grande diffusion du standard. De plus, faire passer les données d'un format à un autre entraîne parfois une perte d'information, dans la mesure l'un des formats peut être moins détaillé que l'autre ; c'est par exemple le cas des informations sur les producteurs d'un jeu de données, qui ne rentrent pas dans les termes du Darwin Core et qui sont généralement exprimées dans un fichier EML accolé aux occurrences décrites en DwC. La perte d'information est alors minimisée pourvu que les deux fichiers restent toujours reliés l'un à l'autre, mais il faut réussir à conserver ce lien dans le temps : les métadonnées sont autant importantes, de notre point de vue, que les données qu'elles documentent.

III. L'ARCHIVAGE DES DONNÉES DE LA BIODIVERSITÉ, ENTRE DONNÉES PHYSIQUES ET NUMÉRIQUES

Le spécimen, numérique ou non, étendu ou non, fait partie des données produites par les chercheurs pour appuyer leurs propos scientifiques. Ces données font partie de ce qu'on appelle les données de la recherche, dans le sens où elles sont produites par les savants dans le cadre de leur activité professionnelle. Cette formulation les rapproche également des archives telles qu'elles sont définies par le Code du Patrimoine à l'article L211-1 :

> « Les archives sont l'ensemble des documents, y compris les données, quels que soient leur date, leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité. »

L'archivage de la recherche scientifique est ainsi possible (voire une obligation légale) dans la mesure où la recherche, en France, est constituée sur fonds publics, et ce d'autant plus que les financeurs nationaux et européens accordent désormais, depuis quelques années, une place importante aux données et aux plans de gestion de ces dernières. Nous étudierons ici les enjeux de la conservation des données de la biodiversité, en gardant la taxonomie comme fil rouge, avant de voir quels sont les moyens pratiques et conceptuels pour ce faire, puis les difficultés pour y parvenir.

A) Enjeux

Les enjeux pour la conservation des données de la biodiversité, et plus particulièrement celles de la taxonomie, sont nombreux. Outre que ces données font partie, dans une certaine mesure, des données publiques et doivent donc répondre à certaines exigences en matière de conservation sur le long terme, elles permettent de mieux appréhender le monde qui nous entoure. C'est de ce dernier point que nous aimerions partir pour élargir à l'utilité de ces données sur le long terme puis à leur hétérogénéité.

1) L'inventaire du vivant en voie de disparition

Les taxonomistes se sont donné pour objectif d'inventorier le vivant, pour mieux connaître les espèces qui forment notre environnement. Au départ pensée pour connaître le comportement des animaux et les propriétés des plantes, l'observation scientifique de la biodiversité s'est doublée, au fil du temps, d'une dimension de contrôle des populations animales et végétales. Dans ce cadre, la description d'une espèce est ainsi assortie du lieu où elle a été observée et du nombre d'individus qui la compose. Certaines espèces, considérées comme nuisibles, ont été (et sont parfois toujours) chassées pour assurer la protection des humains, de ses cultures et de ses élevages ; d'autres ont été surexploitées jusqu'à l'extinction, l'exemple le plus connu étant certainement celui du dodo de l'île

Maurice à la fin du XVII[e] siècle. Cette problématique de la surexploitation et de la disparition des espèces se pose de nos jours en des termes bien plus forts, tandis que que le changement climatique bouleverse les habitats et les modes de vies des populations animales et végétales. Selon une étude scientifique (Larsen et al., 2017), il existerait jusqu'à 6 milliards d'espèces au sein du vivant (micro-organismes compris) et seulement un peu plus d'un million et demi d'entre elles ont été décrites ; ces chiffres sont d'autant plus alarmants que l'on peut considérer que nous traversons actuellement la sixième extinction en masse de la biodiversité. En effet, la vitesse avec laquelle se produisent ces disparitions rend plus cruciale encore la description d'espèces encore inconnues : par exemple, pour estimer les effectifs d'une population, il faut que celle-ci soit délimitée dans une certaine mesure. Même si les exigences de conservation s'incarnent dans des politiques du chiffre qui peinent parfois à saisir la réalité dans toutes ses nuances (Guimont & Petitimbert, 2017), elles ne peuvent s'affranchir de la définition de ce qu'elles comptent. Cette définition, la taxonomie la propose en des termes scientifiques, validés par une communauté d'experts qui s'accordent sur des traits communs à une espèce. Le consensus scientifique est alors la pierre angulaire de la définition d'une espèce.

Cependant, l'état de nos connaissances sur la biodiversité est plus ou moins approfondi, selon les lieux et l'intérêt porté aux taxons. Certaines zones géographiques sont ainsi très bien décrites, d'un point de vue taxonomique, tandis que d'autres restent à explorer plus en détail. Dans des lieux reculés et difficiles d'accès, l'inventaire de la biodiversité est plus délicat, d'autant plus que les grandes expéditions naturalistes ont connu leur âge d'or au cours du XIXe siècle. Elles continuent toutefois de nos jours, parce qu'elles sont essentielles : on parle cependant davantage de « missions » auxquelles participent un ou deux taxonomistes qui sont envoyés sur le terrain. Cela leur permet d'avoir directement sous les yeux leur objet d'étude, et de vérifier les hypothèses qu'ils ont émises depuis leur laboratoire. Les grandes expéditions n'ont pourtant pas tout à fait disparu, comme en témoigne le programme « La planète revisitée »[17], mené par le Muséum National d'Histoire Naturelle, qui enchaîne depuis 2005 les expéditions scientifiques : le Vanuatu, le Mozambique et Madagascar, la Papouasie-Nouvelle-Guinée, la Nouvelle Calédonie, la Guyane et la Corse ont fait l'objet d'inventaires des espèces présentes sur certaines zones de leur territoire. Ces inventaires se concentrent sur des familles qui sont soit méconnues soit qui possèdent de nombreuses espèces en leur sein.

Il faut toutefois, à notre avis, replacer les premières expéditions naturalistes dans le contexte de la colonisation, contexte qui a certes évolué depuis le XIXe siècle. Les expéditions scientifiques d'envergure sont, pour la plupart, financées par les pays occidentaux et ont eu du mal à inclure dans leurs rangs les spécialistes locaux, puis à reconnaître la valeur scientifique de leur travail, alors qu'ils sont pourtant plus à même de connaître la faune et la flore locale. Ainsi, quand bien même les découvertes taxonomiques avaient lieu dans des pays non-occidentaux, elles profitaient davantage aux pays du Nord qui avaient les possibilités (financières, techniques, industrielles) de les exploiter. Pour pallier ce problème évident de redistribution et d'exploitation des ressources et des connaissances, le protocole de Nagoya[18], signé en 2010 et entré en vigueur en 2014, impose le partage des revenus issus de l'exploitation de la biodiversité entre les exploitants et le pays où se situent les ressources exploitées. Cela rend toutefois plus lourde l'organisation de missions et d'expéditions scientifiques, ce qui n'est pas

[17] *La Planète Revisitée.* (s. d.). Muséum national d'Histoire naturelle. Consulté 16 août 2023, à l'adresse https://www.mnhn.fr/fr/la-planete-revisitee

[18] Disponible en français à cette adresse : https://www.cbd.int/abs/doc/protocol/nagoya-protocol-fr.pdf

sans poser quelques problèmes dans une situation où la rapidité est perçue comme essentielle pour contrer les conséquences du réchauffement climatique.

La pratique de l'échantillonnage physique, nécessaire pour l'établissement d'un spécimen holotype, pose question (voir à ce sujet (Rocha et al., 2014) et (Minteer et al., 2014) par exemple). En effet, prélever un individu n'a pas le même impact selon qu'il appartient à une population de vingt, mille ou dix mille individus. La rareté est alors un paramètre à double tranchant : plus l'espèce est rare, plus il est nécessaire de la documenter avant qu'elle ne disparaisse, mais en prélever un individu risque alors d'accélérer cette extinction. C'est dans ce contexte que le spécimen numérique est apparu comme une solution idéale pour les non-taxonomistes : il aurait permis, presque mécaniquement dans le cas du séquençage ADN, de délimiter les espèces grâce à des prélèvements non-intrusifs. Dans le cadre du spécimen numérique étendu, pouvoir conserver sur le long terme des données qui servent à l'identification d'espèces est essentiel. S'appuyer uniquement sur les technologies numériques, c'est toutefois faire reposer une science non sur un savoir mais sur une unique pratique, et minimiser ainsi drastiquement le travail des taxonomistes. Que le spécimen soit numérique ou non, le conserver permet de mieux connaître le monde qui nous entoure, et, dans le cas de spécimen holotype, il sert de référence pour une espèce.

2) La très longue durée de vie des données de la taxonomie

Par ailleurs, notons que les données de la taxonomie peuvent être mobilisées sur le très long terme. En effet, la description d'une espèce est parfois utilisée pendant de nombreuses décennies, même si elle est revue et augmentée au fil du temps.

Or, la référence à une espèce dans le cadre d'un article taxonomique inclut diverses informations : en plus du nom latin du taxon, il est commun d'ajouter le nom de la personne a avoir proposé l'hypothèse taxonomique et la date de la description du spécimen : c'est ainsi que Linné (sous l'abréviation « L. ») ou Cuvier apparaissent encore dans les publications scientifiques actuelles. Ces deux informations sont essentielles car elles permettent de citer avec précision une version de l'état des connaissances à un instant particulier. Cela permet également de s'inscrire dans une filiation scientifique et de rendre crédit à la personne qui a décrit (ou précisé, selon le cas) le taxon. Sur le long terme, il faudrait ainsi conserver ces données, et d'autres, à fonction probante. Ces dernières ont servi à établir l'espèce, et font d'ailleurs parfois partie du spécimen numérique étendu : analyse de la salive d'un escargot, reconstitution 3D de sa coquille, photographie de son habitat, séquençage ADN, positionnement dans l'arbre phylogénétique, distance avec les plus proches voisins dans celui-ci… En plus des données numériques, il faut également prendre en compte, dans cette masse d'éléments probants, les échantillons qui ont été prélevés sur l'individu, ainsi que celui-ci, pour déterminer une espèce : tout ce faisceau de preuves sert à démontrer la cohérence d'un taxon et doit donc être conservé sur le long terme. Notons par ailleurs que l'arrivée de nouvelles technologies peut permettre de nouvelles découvertes sur des spécimens anciens : c'est par exemple le cas du séquençage ADN de tissus conservés dans du formol, qui ne pouvait se faire il y a quelques années (Hahn et al., 2022).

Du fait de leur statut de référence, les spécimens holotypes nécessitent une conservation sur le long terme, conservation qui s'étend également aux données qui forment le spécimen numérique étendu qui y est relié. Il s'agit alors de pouvoir conserver aussi bien la monographie ou l'article qui décrit le taxon et ses caractéristiques que ce qui a permis la découverte, et donc les données rattachées à l'article.

De plus, conserver un spécimen ou les données qui s'y rattachent, c'est également faciliter la recherche scientifique dans la mesure où il est possible que ces données soient réutilisées par d'autres scientifiques, à plus ou moins long terme. Les spécimens sauvegardent en quelque sorte un état de l'espèce à un instant donné, et disposer de données au sujet d'un taxon sur le long terme permet d'en retracer les évolutions. Posséder des données sur plusieurs individus répartis dans le temps permet de les comparer entre elles et possiblement d'ouvrir des pistes de recherche ; ces comparaisons à grande échelle sont également utiles, d'un point de vue statistique, pour déterminer les moyennes des caractéristiques d'une espèce.

Les enjeux de conservation des données de la taxonomie, et de la biodiversité en général, rejoignent ainsi ceux des autres données de la recherche : il s'agit de les conserver pour les utiliser comme preuves scientifiques de ce que l'on avance dans un article scientifique et d'en faciliter la réutilisation. C'est également une étape essentielle à la FAIRisation des données de la recherche : l'archivage, par ses procédés d'indexation et de préservation de l'accès dans le temps aux données, permet de cocher les principes de « facile à trouver » et d'« accessibilité », au moins.

Des efforts existent déjà pour redonner accès aux descriptions morphologiques anciennes : la *Biodiversity Heritage Library* met ainsi en ligne des numérisations d'ouvrages anciens pour permettre d'accéder aux anciennes descriptions scientifiques. Soixante millions de pages sont disponibles à l'heure actuelle[19], représentant une véritable mine d'informations pour les scientifiques, dans la mesure où ces textes anciens ne sont pas facilement accessibles.

De façon plus large, sauvegarder les données de la biodiversité est important dans le contexte environnemental qui est le nôtre, à savoir le changement climatique et les changements qu'il apporte. À l'heure de la disparition effrénée des espèces végétales et animales, documenter au mieux le vivant et ses conditions d'épanouissement permet non seulement de transmettre aux générations futures un état des lieux mais surtout d'agir de façon éclairée pour leur conservation.

3) L'hétérogénéité des données de la taxonomie

De par ses pratiques, la taxonomie mobilise divers types de données et de documents, aussi bien numériques que physiques.

En effet, la morphologie utilise les spécimens physiques, conservés pour certains dans de l'éthanol ou du formol, pour d'autres, sous forme naturalisée ; il est également possible de faire appel aux squelettes des animaux, en plus de l'aspect extérieur de ceux-ci. Du côté des végétaux, le spécimen peut s'incarner de différentes manières : la fleur, les feuilles, les graines, la tige… Pour les microbes et bactéries, on parle parfois de « collections vivantes » car les souches sont maintenues en vie, pour être diffusées auprès de laboratoires de recherche et d'universités, dans l'optique de faciliter leur étude scientifique. De façon générale, tout ce qui forme le spécimen est susceptible de faire l'objet d'une conservation dans les collections naturalistes, et les supports de ces

[19] *Biodiversity Heritage Library*. (s. d.). Consulté 24 août 2023, à l'adresse https://www.biodiversitylibrary.org/

spécimens sont variés : les pages d'un herbier, par exemple, nécessitent un autre conditionnement physique qu'un bocal de formol, par exemple, ou qu'une planche d'insectes… Ces dernières sont d'ailleurs fragiles, et requièrent des précautions à l'usage, surtout si elles accusent un certain âge. Pour ce qui est des animaux naturalisés, s'ils présentent l'animal dans un état qui se veut fidèle à sa vie dans la nature, il existe toutefois des conditions optimales pour leur préservation.

Notons également que les données de la taxonomie incluent également les notes prises par les chercheurs ainsi que leurs publications scientifiques, qui peuvent exister aussi bien au format papier que numérique. Il peut aussi exister des différences majeures ou mineures entre ces deux versions, ce qui pose la question de la version à archiver, tandis que les capacités de stockage sont limitées.

De plus, certaines données n'existent qu'au format numérique : il s'agit par exemple des résultats des séquençages ADN, mais aussi des tableurs de suivi de la population d'une espèce, ou bien de ceux contenant les caractéristiques de différents spécimens collectés lors d'un inventaire ou d'une expédition scientifique. Ces derniers, du moins s'ils suivent le formalisme du Darwin Core, sont accompagnés d'un autre document qui précise qui a participé à la collecte des informations et les affiliations universitaires de chacune de ces personnes. De plus, n'oublions pas que les données qui forment le spécimen numérique étendu sont également à compter parmi les données de la taxonomie : enregistrements du barrissement d'un éléphant, taille et forme de ses empreintes, noms des plantes qu'il consomme, ses parasites, mais aussi son comportement avec ses congénères et, plus globalement, ses interactions avec son environnement.

Par ailleurs, tandis que certaines des données mentionnées plus haut peuvent être mobilisées par d'autres disciplines au sein de la biologie, certaines données sont spécifiques à la taxonomie. L'article « *Repositories for Taxonomix Data : Where we are and what is missing* » (Miralles et al., 2020) différencie une vingtaine de types de données, dont de nombreuses possèdent des extensions de fichiers différentes : si on y retrouve beaucoup de formats ouverts (comme le .csv) et recommandés par les archivistes pour la conservation sur le long terme (comme le .tiff), d'autres ne sont pas acceptés par les bases de données en ligne (comme pour la spectroscopie infra-rouge) et sept nécessiteraient, selon les auteurs de l'article, le développement de bases et de standards adéquats. Ces données sont alors éparpillées, selon leur type, dans différentes bases de données, lorsqu'elles existent.

Prenons ainsi l'exemple du genre *Rosa L.* tel qu'on peut le télécharger sur la Catalog of Life Checklist[20]. Différents formats sont possibles pour notre jeu de données : tout d'abord une archive Darwin Core, puis les formats préférés par le CoL (à savoir l'ACEF et le COLDP, tous deux développés pour la Checklist précisément, le second venant remplacer le premier à partir de 2014), et enfin trois formats pour la représentation graphique des arbres (textree, newick et dot). Choisissons le COLDP pour son exhaustivité : il possède en effet plusieurs fichiers tsv détaillant la répartition géographique des taxons, des liens vers des médias pour chacun d'eux, les relations entre les différents noms associés à chacun des taxons, les publications décrivant les taxons pour la première fois, le nombre d'espèces qui constituent le taxon, les interactions entre ces taxons et d'autres, les relations entre

[20]Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., Stjernegaard Jeppesen, T., Örn, A., Vandepitte, L., Hobern, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Abbott, J., Adlard, R., Adriaenssens, E. M., et al. (2023). *Catalogue of Life Checklist* (Version 2023-07-18) [Dataset]. Catalogue of Life. https://doi.org/10.48580/dfsy

les taxons, le type de documents reliés, et enfin les noms vernaculaires. Pour relier tous ces fichiers, un identifiant est associé à chacun des taxons et est rappelé en début de ligne ; notons que d'autres fichiers, aux formats respectivement YAML et JSON, donnent pour le premier la liste des personnes ayant participé à l'élaboration du CoL et les références compatible au format CLS-JSON.

L'utilisation de ces différents fichiers ne peut se faire que si l'utilisateur est déjà familier avec les principes de la taxonomie, mais l'ensemble est pensé pour être facilement lisible sur plusieurs supports : le format tsv est ouvert et facile d'utilisation, tout en étant moins répandu que le csv.

Conserver les données de la taxonomie, c'est ainsi garder une trace de notre époque et de sa diversité biologique pour les générations futures. En conservant les données dans leur état actuel, nous fournissons aussi des pistes pour les historiens des sciences pour enquêter les pratiques actuelles. Enfin et surtout, c'est fournir aux biologistes des fondations solides pour leurs travaux de recherche.

B) Moyens pratiques et conceptuels

1) Les données de la recherche, entre injonctions à la science ouverte et archivage

La science ouverte est un mouvement qui gagne en puissance depuis une trentaine d'années : au début, elle concernait principalement l'accès aux articles scientifiques, dont le coût pour les chercheurs était un véritable frein à la recherche scientifique. Mais depuis quelques années, le mouvement s'intéresse de plus en plus aux données de la recherche, pour les rendre accessibles à tous, afin de favoriser la reproductibilité des résultats de recherche et leur réutilisation.

Entre France, cette notion s'appuie sur les différents plans pour la science ouverte, ainsi que sur les lois Valter (2015) et Lemaire (2016) qui postulent l'ouverture des données produites par les scientifiques, surtout si elles ont été produites dans le cadre de projets de recherches financés au moins à moitié sur fonds publics. La section Aurore de l'AAF définissait en 2014 les données de la recherche comme « […] l'ensemble des informations et matériaux produits et reçus par des équipes de recherche et des chercheurs. Elles sont collectées et documentées à des fins de recherche scientifique. À ce titre, elles constituent une partie des archives de la recherche. »[21]. Cette définition fait appel au concept des archives de la recherche, que l'on peut rapprocher de l'expression « archives scientifiques » ou « archives des sciences » (Charmasson, 2006) : effectuer ce rapprochement permet ainsi de faire tomber les données de la recherche dans le cadre légal de l'archivage public en France.

Au-delà des obligations législatives, l'archivage des données de la recherche est important dans la mesure où il permet le partage d'une version définitive d'un jeu de données. Conserver les données de la science, de façon plus générale, c'est sauvegarder l'état de nos connaissances à un moment donné : c'est également permettre aux chercheurs de diffuser leurs connaissances entre eux, de s'inscrire dans des communautés de recherche, et de faciliter leur intégration dans le monde scientifique en

[21]Pomart, J. (2014, juillet 18). AAF / Section Aurore : Un groupe de travail sur les données de la recherche [Billet]. *Archives de la FMSH*. https://archivesfmsh.hypotheses.org/1209

général, où la citation est le moyen de se faire remarquer, et donc d'obtenir des postes et des financements pour des projets de recherche (Farnham et al., 2017). La science ouverte fournit le cadre intellectuel à l'ouverture des données de la recherche ; et le respect des principes FAIR permet d'ouvrir ces données, tout en gardant en tête les contraintes qui pèsent sur leur diffusion. La formule « Aussi ouvertes que possible, aussi fermées que nécessaire » rend bien compte de cette ambiguïté : tout n'est pas diffusable à grande échelle, pour des questions aussi bien éthiques que financières, mais le partage de ce qui peut être rendu disponible permet à la fois de justifier des résultats publiés et de transmettre ces connaissances à d'autres personnes qui pourront les réutiliser dans d'autres contextes.

Dans le contexte de la taxonomie et de la biologie en général, les données de la recherche contiennent par exemple aussi bien les spécimens que les informations conservées à leur sujet, parmi lesquelles on trouve les séquençages ADN, les photographies d'individus, leurs localisations ; on peut également y compter les logiciels développés pour l'analyse automatique des pas entre les différentes branches d'un cladogramme, par exemple, et les résultats de ces analyses.

On voit également apparaître depuis quelques années la revendication d'un spécimen ouvert (Colella et al., 2021). Ce dernier promeut la diffusion des spécimens conservés dans les collections naturalistes, et plaide pour leur intégration aux plans de gestion des données réclamés par les financeurs lors du montage d'un projet de recherche, sur le constat suivant : ce sont les données « secondaires », dérivées des individus collectés (comme par exemple les analyses ADN ou les photographies), qui sont le plus souvent concernées par les plans de gestion de données, tandis que les spécimens, qui sont pourtant la source de ces données secondaires, sont très rarement mentionnés. Intégrer les spécimens dans les plans de gestion permettrait de rappeler leur importance pour la recherche scientifique : du fait de leur unicité et de leur valeur pour la recherche en biologie, et comme source de données dérivées, penser leur conservation et leurs réutilisations possibles peut se faire dès la collecte, voire avant. En les intégrant dans ces plans de gestion des données, on met également en avant la question du financement des collections naturalistes, question cruciale pour la conservation de nos connaissances sur la biodiversité sur le long terme.

Toutefois, partager les données de la recherche ne peut se faire sans une réflexion pour les rendre réellement disponibles à l'intégralité du monde de la recherche, au-delà des questions techniques. Un compte-rendu de conférence intitulé « *Open science, data sharing and solidarity: who benefits?* » (Staunton et al., 2021) met ainsi en évidence le fait que les données de recherche sont le plus souvent produites par les pays développés, selon leur propre agenda, et selon leurs propres conditions, qui sont parfois défavorables aux autres pays qui ne possèdent pas encore les capacités techniques d'accès aux données déposées, ni de les réutiliser, ni en récupérer les bénéfices financiers. Dans cette optique, les conventions internationales comme la Convention sur la Diversité Biologique de 1992 ou le protocole de Nagoya (2010) offrent un cadre juridique contraignant les réutilisations, mais sont jugés insuffisants par les auteurs de l'article.

La particularité des données de la recherche dans la biologie est qu'elles recouvrent une large gamme de formats (voir sous-partie précédente) : de ce constat, il en ressort que les procédures d'archivage doivent être adaptées au type

des données considérées. Nous étudierons ainsi d'abord les données numériques et les spécimens et échantillons qui sont conservés par les collections naturalistes.

2) Les bases de données de la biodiversité

Pour conserver sur le long terme les données numériques de la biodiversité, l'une des premières solutions auxquelles on peut penser concerne les bases de données – taxonomiques ou non – disponibles en ligne. Elles sont de différentes catégories, que nous allons détailler ici.

Il existe ainsi des bases de données spécialisées dans un type de données : c'est par exemple le cas de GenBank[22] ou de l'ENA[23] (*European Nucleotide Archive*) qui stockent toutes les deux des séquençages génétiques. Ces dernières sont disponibles, le plus souvent, sous deux formats : FASTA (défini par GenBank) et l'EMBL (pour l'ENA), sous forme de fichiers textes suivant des formalismes pré-définis, assortis d'un certain nombre de métadonnées. Notons d'ailleurs que ces dernières sont beaucoup plus nombreuses dans le cas de l'EMBL (entre autres : noms des auteurs, références des publications, placement de l'espèce dans la classification linnéenne, date, somme de contrôle MD5, gène visé, et enfin séquence ADN en elle-même) que pour le format FASTA (identifiant, nom scientifique de l'espèce, type de séquençage, séquence).

Un autre type de bases de données concerne les listes d'espèces, comme par exemple le CoL (*Catalogue of Life*)[24], qui a pour objectif de réunir l'ensemble des taxons connus dans une liste accessible à tous, afin d'obtenir un référentiel des taxons utilisés et valides d'un point de vue taxonomique. Ces référentiels permettent de replacer avec précision un taxon dans la classification linnéenne et de connaître les différents noms qui ont été rattachés au taxon recherché. Pour ce faire, CoL pioche dans différentes bases présentes en ligne, et qui se veulent également des listes d'espèces dans un domaine précis de la biodiversité : parmi les sources du CoL, on peut penser à WoRMS (*World Register of Marine Species*), l'ITIS (*Integrated Taxonomic Information System*) ou encore l'Index Fungorum, pour n'en citer que trois sur les 163 bases moissonnées par le CoL[25]. Lorsque l'on sélectionne un taxon, apparaissent d'abord son identifiant unique au sein du CoL, puis son nom scientifique, sa place dans la classification linnéenne, le nom de sous-éléments que possède ce taxon, le ou les noms vernaculaires, les références dans la littérature scientifique ; les trois dernières catégories concernent la ressource initiale qui a permis la création de la notice par le CoL, à savoir le jeu de données dont les informations sont issues et un lien vers la ressource originale.

Un troisième type de plateforme auquel on peut se retrouver confronté est le super-portail, qui permet par une unique requête de fouiller plusieurs bases à la fois : c'est par exemple le cas du GBIF[26], qui propose sur son site internet de récupérer des données issues de bases précédemment mentionnées. On peut ainsi fouiller, depuis la même interface, la CoL Checklist (c'est-à-dire la version accessible par API du CoL), l'ENA et des jeux de données issus d'applications mobiles (comme eBird, iNaturalist ou PlantNet), en plus des jeux de données déposés directement par les chercheurs qui décrivent leurs observations scientifiques. La recherche d'un taxon permet de visualiser plusieurs types de résultats lorsqu'on clique sur le nom de l'espèce : d'abord le nom

[22] *Genbank overview*. (s. d.). Consulté 20 août 2023, à l'adresse https://www.ncbi.nlm.nih.gov/genbank/

[23] *Ena browser.* (s. d.). Consulté 20 août 2023, à l'adresse https://www.ebi.ac.uk/ena/browser/home

[24] *Col*. (s. d.). COL. Consulté 20 août 2023, à l'adresse https://www.catalogueoflife.org/

[25] Chiffres disponibles sur *Source datasets*. (s. d.). COL. Consulté 20 août 2023, à l'adresse https://www.catalogueoflife.org/data/source-datasets

[26] *Gbif*. (s. d.). Consulté 23 août 2023, à l'adresse https://www.gbif.org/

scientifique et sa publication, ainsi que la source de ces informations, le nom vernaculaire et éventuellement le basionyme. Mais surtout, on peut consulter les différents médias attachés aux occurrences d'observation de cette espèce (images, enregistrements sonores) ainsi que la géolocalisation des observations ; il est également possible d'accéder à tous types d'informations concernant l'espèce, comme par exemple là où est conservé le spécimen holotype, la description de l'habitat, son aire de répartition, le risque d'extinction selon l'IUCN (*International Union for Conservation of Nature*) etc. On retrouve ainsi sur ce portail ce qui, à notre avis, se rapproche le plus du spécimen numérique étendu : il manque ainsi principalement les liens avec les autres espèces (notamment les parasites) pour disposer d'un aperçu complet de l'espèce consultée. Notons toutefois que les données présentées font références à des occurrences hébergées dans d'autres bases de données, et donc à d'autres spécimens que le spécimen holotype. Aujourd'hui, le GBIF est la principale porte d'accès aux données de la biodiversité, de par son moissonnage de nombreuses bases et sa compatibilité technique : le Darwin Core sert de standard commun à l'ensemble des bases accessibles via la plateforme du GBIF.

Enfin, on peut retrouver les données de la recherche en biologie sur les entrepôts de données, institutionnels ou non, spécialisés par discipline ou non : par exemple Dryad[27] permet de mettre en ligne un jeu de données relié à un article sans contrainte de forme. Le chercheur y dépose ses fichiers à l'intérieur d'un dossier zippé et reçoit en échange un DOI ainsi qu'un lien pérenne qui pointe vers ces fichiers. Chacun peut ainsi déposer ses données, mais celles-ci ne sont pas moissonnées par les bases spécialisées et donc se retrouvent éparpillés parmi d'autres données, issues d'autres disciplines.

Néanmoins, nous aimerions attirer l'attention sur le fait que les bases de données ne présentent pas toutes les caractéristiques de l'archivage de données numériques. Le respect des principes FAIR, qui mènent entre autres à la conservation des données sur le long terme, est un premier pas ; mais il faut, à notre avis, distinguer la mise en ligne de l'archivage pérenne de données numériques. Ainsi, même si le terme d' « archive » pour parler des bases de données en ligne est de plus en plus couramment utilisé par d'autres professions que les archivistes, il ne faut pas négliger l'ensemble des traitements à effectuer pour que l'on puisse considérer une donnée comme véritablement archivée. L'attribution d'identifiants uniques est ainsi considérée et fortement recommandée (Güntsch et al., 2018; Poisot et al., 2013), de même que l'adoption de standards communs tels que le Darwin Core ou l'ABCD ; en dehors de ces questions, la réplication des données est souvent perçue par les chercheurs comme la seule façon d'assurer la conservation des données sur le long terme (voir par exemple l'appendice 12 de l'article « *Repositories for Taxonomic Data: Where We Are and What is Missing* » (Miralles et al., 2020)) : c'est toutefois négliger les procédures de vérification d'intégrité des données par la vérification automatisée des sommes de contrôle des fichiers, ou les normes internationales relatives à l'archivage électronique, comme par exemple le modèle OAIS (ISO 14721) et les standards internationaux de métadonnées pour l'archivage (comme METS, par exemple). Le temps nous manque pour mener cette étude de façon plus approfondie : les premières pistes d'études seraient par exemple du côté de l'application de ces normes par les bases de données.

[27] *Dryad | Home—Publish and preserve your data*. (s. d.). Dryad. Consulté 20 août 2023, à l'adresse https://datadryad.org/stash

3) Les collections d'histoire naturelle

Les muséums d'histoire naturelle ont derrière eux un long historique de conservation des données primaires de la biodiversité : leurs collections sont constituées de spécimens collectés au fur et à mesure des expéditions scientifiques au cours des siècles. Aujourd'hui, elles comptent plus d'un milliard de spécimens, principalement concentrés dans l'hémisphère nord et les pays occidentaux (Johnson et al., 2023), qui sont tous des ressources essentielles à la taxonomie mais aussi à l'ensemble de la biologie. L'étude de ces collections permet d'obtenir une vue d'ensemble du vivant, mais aussi de l'état des connaissances, à une époque donnée. Conserver ces documents uniques permet ainsi de connaître l'état passé de la biodiversité, ce qui nous éclaire sur son état actuel.

Aujourd'hui, les scientifiques semblent redécouvrir les collections d'histoire naturelle, dans la mesure où les objets qui y sont conservés sont exploitables d'une nouvelle façon : il est maintenant possible d'obtenir le séquençage ADN de spécimens conditionnés il y a plus de cent ans (Raxworthy & Smith, 2021), même si cela pose parfois des problèmes lorsque les spécimens ont été mal identifiés à l'origine et que le dépôt en ligne de ces séquençages se fait sans corriger les erreurs initiales (Mulcahy et al., 2022). De nouvelles applications semblent ainsi se dessiner pour les collections d'histoire naturelle : l'automatisation de l'identification de plantes grâce à l'intelligence artificielle (Mora-Cross et al., 2022) ou la comparaison de spécimens collectés aux mêmes endroits sur plusieurs années, ce qui en font des ressources particulièrement adéquates pour étudier les effets du changement climatique sur une longue échelle temporelle (Lister, 2011), en sont deux exemples parmi d'autres.

Ceci est possible grâce à la numérisation des collections, qui permet d'accéder à des données alors peu connues : mais la mise en ligne des collections naturalistes n'est pas encore complète (pour un exemple en paléontologie, voir par exemple « *Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution* » (Marshall et al., 2018)).

Le rôle des collections naturalistes au XXIe siècle est ainsi plus large qu'initialement pensé : comme l'écrit Roselli Pellens dans son ouvrage *Les collections naturalistes dans la science du XXIe siècle: une ressource durable pour la science ouverte* (p.12) :

> « Rétrospectivement, il aurait été difficile à la fin du XIXe siècle d'imaginer la plupart des recherches actuellement menées sur les collections. Il en va de même de notre capacité à prévoir en détails les besoins sociétaux et les regards scientifiques pour le prochain siècle ou le suivant. »

Ainsi, il nous semble important de penser au futur des collections naturalistes telles qu'elles existent actuellement. Avec la sixième extinction de masse actuellement en cours, les collections et les spécimens qu'elles abritent sont d'autant plus cruciaux qu'il sera peut-être possible de les mobiliser de façon inattendues par rapport à nos connaissances actuelles. La collecte de spécimens et la documentation afférente devrait ainsi être archivée sur le long terme, afin que les informations puissent être mobilisées à travers le temps : notons que de nos jours les catalogues de collections sont informatisés, et que, si leur mise en ligne au moins partielle permet l'accessibilité des collections qu'ils documentent, ce sont également des bases de données dont la préservation sur le long terme est essentielle pour la connaissance de la biodiversité. Il nous semble important de sauvegarder le plus d'informations pour les années à venir : de façon générale, la mise en ligne des informations présentes dans les collections n'est selon

nous pas satisfaisante si elle n'est pas accompagnée des métadonnées adéquates, techniques et contextuelles à la fois.

S'il n'est pas possible de prévoir les utilisations des collections naturalistes dans le futur, on peut néanmoins s'interroger sur la mise en place de protocoles qui permettront le plus de réutilisations par la suite : cela passe par la connaissance précise des spécimens présents dans les collections, et donc, entre autre, leur identification correcte par les taxonomistes, mais aussi l'historique des identifications attribuées à un spécimen, sa localisation et sa date de collecte précises, le nom du récolteur. Si ces informations sont « classiques », il nous semble important de les conserver autant que faire se peut, car elles permettent de se faire une représentation, *a posteriori*, du territoire sur lesquels les spécimens ont été recueillis. Dans cette perspective, le concept de spécimen étendu permet d'élargir la gamme des informations présentes dans les collections naturalistes : puisque l'individu collecté est replacé au cœur de son environnement, l'ensemble des informations permet de mieux connaître les espèces, y compris sur le long terme. Ces données, il nous semble qu'il faut les collecter au plus tôt, et également les conserver sur le long terme. Recueillir le plus d'informations possible avant leur disparition : voilà un des défis auxquels les biologistes sont confrontés aujourd'hui. Le rôle des collections naturalistes est alors d'assurer la conservation sur le long terme des spécimens mais également des informations associées.

De plus, en dehors du métrage d'étagères et des bocaux de formol, il nous semble qu'il ne faut pas négliger les compétences humaines d'identification des espèces : les taxonomistes, grâce à leur expertise sur les taxons, permettent de vérifier les identifications proposées pour un spécimen donné. Cela nous semble d'autant plus important que les espèces peuvent être reclassées dans la hiérarchie linnéenne ; connaître l'historique des noms, c'est-à-dire posséder les documents pour ce faire, permet de retracer l'histoire des connaissances scientifiques au sujet d'un taxon donné.

Il faut toutefois se poser la question de ce qu'on collecte, puisque les espaces de stockages ne sont pas infinis, aussi bien pour les serveurs que pour les collections naturalistes. La conservation des données de la biodiversité a un coût matériel important, mais il nous semble inférieur aux implications de la disparition de connaissances sur des milliers d'espèces.

Il nous semble important de détailler à présent les différentes difficultés que nous identifions à l'archivage des données de la biodiversité.

C) Les difficultés de l'archivage des données de la biodiversité

Néanmoins, l'archivage des données de la biodiversité ne semble pas faire l'objet d'une véritable attention, aussi bien de la part des archivistes que des biologistes eux-mêmes. Nous aimerions ainsi étudier et réfléchir aux raisons de ce désintérêt que nous semblons remarquer dans la littérature scientifique, aussi bien du point de vue de la biologie que des archivistes. Nous nous interrogerons ensuite sur la place du numérique dans les pratiques des biologistes du point de vue de l'archivage, avant de considérer les premières avancées du domaine.

1) Un faible intérêt de la part des archivistes et des biologistes ?

La figure de l'archiviste, pour les non-professionnels de l'information, est souvent caricaturée en une personne qui s'occupe de documents poussiéreux, dans une petite salle sombre au sous-sol ; il y a une part de vrai dans cette image, dans la mesure où les archives sont encore bien souvent des documents papier, principalement conservés pour leur valeur juridique et administrative, et qui n'acquièrent qu'au fil du temps leur valeur historique. Les données, souvent sous format numérique, ne relèveraient pas du domaine d'expertise des archivistes, qui devraient s'occuper des factures de laboratoire, des documents administratifs, etc. Ainsi, même le concept d'archives scientifiques tel que Thérèse Charmasson le théorise fait la part belle aux documents administratifs, qui sont présentes à la fois dans les « archives de tutelle des établissements de recherche » et des « archives propres de ces mêmes établissements » (Charmasson, 2006). Viennent seulement en troisième temps les archives personnelles des chercheurs, au double statut ambigu, entre documents produits dans le cadre d'une activité professionnelle et perçues comme archives privées (parce que personnelles) par leurs producteurs.

Les données de la recherche, on l'a vu un peu plus tôt, rentrent dans le cadre théorique et juridique des données publiques, du moins lorsque les projets qui ont mené à leur production ont été financés au moins à moitié sur fonds publics. Parce qu'elles sont encore utilisées ou en cours d'utilisation par leurs producteurs, à notre avis, les archivistes ont eu tendance à les considérer comme en dehors de leur compétence : la théorie des trois âges, mise au point par Yves Perrotin au tournant des années 1960, postule en effet que les documents ne deviennent véritablement des archives que lorsqu'ils quittent leur fonction première de preuve et qu'ils peuvent être mobilisés par d'autres personnes que leurs producteurs originaux, et dans des contextes tout à fait différents. Cette conception ne nous semble pas convenir aux données de la recherche, dans la mesure où la diffusion et la réutilisation dans d'autres contextes peut intervenir très rapidement : les données ne passent pas par un « âge intermédiaire » durant lequel elles seraient moins mobilisés sur une longue période par un chercheur, sauf si elles n'ont pas été diffusées au préalable. C'est toutefois de moins en moins le cas : l'ouverture des données est une des incitations qui pèsent fortement sur les scientifiques.

La réutilisation des données de la recherche nous semble participer du troisième âge des archives, puisque les données sont alors remobilisées par d'autres personnes que leurs producteurs, et pour des buts parfois tout à fait différents. Le cycle de vie des données ne pose toutefois pas la conservation des données comme la fin du cycle : les données, une fois déposées dans un entrepôt, peuvent alors être réutilisées par d'autres. Les bibliothèques ont globalement réussi à imposer leur présence auprès des chercheurs dans le dépôt des données (Mesguich, 2023) ; mais les archivistes restent à la marge de ces mouvements, alors qu'on pourrait penser qu'ils seraient intéressés par la conservation sur le long terme de documents qui peuvent acquérir au fil du temps une valeur historique.

Le peu d'intérêt porté par les archivistes aux données de la recherche se retrouve dans la littérature : nous n'avons ainsi pas trouvé de publications à ce sujet, aussi bien dans la littérature francophone qu'anglophone. Le seul résultat marquant est celui d'un dossier consacré par la revue scientifique *Archival Science* publié en mars 2007 (Volume 7, issue 1), consacré à ce sujet. Un autre article intéressant concerne la distinction entre données chaudes et froides, effectuée par Cyril Pernet et ses collègues dans un article paru dans *Neuroinformatics* en 2023 (Pernet et al., 2023). Un dossier a été consacré en 2017 à la question des archives de la recherche par la *Gazette des* archives (n°246, 2017-2), mais ne mentionne pas les données de la recherche. À part cela, nos recherches

bibliographiques ont été infructueuses. Il faut à cela rajouter que les sciences de la nature ne ressortent que très peu dans les quelques articles qui traitent de la question : il faut aller piocher dans les revues de la discipline en question, pour ne trouver à nouveau que quelques articles traitant de l'état des bases de données, et plaidant pour une meilleure gestion des données. L'article « *The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs* » (Akmon et al., 2011) donne toutefois à voir certaines des problématiques étudiées ici : les données sont essentielles mais peu partagées au début des années 2010 et le départ de leur producteur, qui est également le seul à connaître le contexte de leur création, en empêche la réutilisation par les autres membres du laboratoire. Si les chercheurs interrogés dans cet article semblent conscients de l'importance de l'accès aux données, ils ne paraissent pas exactement comment le mettre en place de façon effective pour tous.

Un autre exemple de cette tendance se retrouve dans une conférence donnée lors des journées TDWG 2022 intitulée « The Importance of Collecting and Archiving Data on Domestic and Cultivated Organisms » par Quentin Groom (Groom et al., 2022) : ce dernier argumente pendant sa présentation sur l'importance de connaître les interactions entre espèces et finit sa conférence sur un « mettez en ligne vos données » mais ne parle pas d'archivage de la façon dont les archivistes le conçoivent.

Nous avons également mené deux entretiens axés, entre autres, sur la question de la préservation des données de la recherche et l'importance que les chercheurs et pouvaient accorder à cette question.

D'après l'entretien mené avec Mme Sophie Pamerlon, ingénieure données au GBIF depuis une dizaine d'années, les bases de données n'ont pas vocation à l'archivage pérenne, mais une fonction de stockage des données à moyen terme ; toujours selon elle, les chercheurs ne sont pas toujours sensibles à l'archivage de leurs données. La science ouverte et son injonction au dépôt des données permettent de « sauvegarder », dans une certaine mesure, les données des chercheurs : il s'agit davantage de stockage que d'archivage pérenne.

Un autre entretien, mené avec Mme Roselli Pellens, macroécologiste à l'Institut de Systématique, Évolution, Biodiversité (UMR 7205 – CNRS – MNHN – EPHE-PSL – Université des Antilles) confirme cette tendance : l'archivage est envisagé par la chercheuse comme la réplication des données sur des supports différents, et non comme la conservation à long terme des données produites. Elle est néanmoins consciente de l'importance de la conservation des spécimens sur le long terme, et elle insiste sur le sujet dans les formations qu'elle donne. Elle dit également insister sur la provenance des données et estime que la sensibilisation aux métadonnées est essentielle auprès des producteurs de données, du fait de la variété des pratiques : il existe ainsi plusieurs façons de rentrer un nom scientifique dans un tableur (en deux champs séparés ou un seul, avec des abréviations ou non, etc.).

Il nous paraît pertinent d'analyser les entretiens au regard de l'impensé informatique, tel que conceptualisé par Pascal Robert (Robert, 2010) page 381 de son ouvrage *Mnémotechnologies.* Ce concept postule que la vision actuelle des outils numériques est si positive qu'elle en empêche toute réflexion critique sur leur usage ; à notre avis, cet impensé numérique peut également s'étendre à la production et à l'archivage des données, produites dans un cadre scientifique ou non. Ainsi, la conservation des données de la recherche, de façon plus générale, ne

s'impose pas d'emblée aux producteurs desdites données, puisqu'elles ne sont alors pas perçues comme des documents nécessitant d'être archivés. Ainsi, même un document qui se veut une présentation synthétique des *bioinformatics*, mélange constamment la publication et la préservation des données (Gadelha et al., 2021).

2) Une certaine réticence aux technologies numériques ?

Les technologies numériques sont le point de mire des financeurs et des biologistes de façon générale, mais la taxonomie ne peut, du fait de ses objets d'études, concevoir un passage au tout numérique. Les spécimens physiques sont en effet la source des données primaires de la taxonomie : c'est sur eux que l'on se fonde pour les analyses secondaires que sont par exemple les divers séquençages ADN possibles, les comparaisons morphologiques, etc.

La littérature taxonomique insiste en effet sur l'importance du spécimen physique pour la vérification sur le long terme des identifications ainsi que pour les différentes réutilisations possibles du fait des progrès technologiques. Ne pas déconnecter les spécimens des données primaires de la biologie, pour reprendre le titre de l'article de Julien Troudet et de ses collègues (Troudet et al., 2018), est alors essentiel pour permettre les réévaluations d'identification, et la collecte de données secondaires dans un second temps.

De plus, d'après un entretien mené avec M. Aurélien Mirallès, taxonomiste à l'Institut de Systématique, Évolution, Biodiversité (UMR 7205 – CNRS – MNHN – EPHE-PSL – Université des Antilles), certains taxons, comme par exemple chez les insectes, ne sont pas encore suffisamment étudiés pour que le recours aux analyses ADN soit nécessaire pour reconnaître qu'une espèce n'a pas encore été décrite par la communauté scientifique. M. Mirallès note toutefois que cela n'est pas le cas pour son domaine d'expertise, à savoir les amphibiens, où les indices qui permettent désormais de reconnaître une espèce comme nouvelle sont de plus en plus subtils, ce qui rend pertinent l'usage d'un séquençage ADN pour distinguer les taxons entre eux.

À la lumière du *taxonomic impediment*, il nous semble toutefois que c'est la façon dont les nouvelles technologies ont été amenées dans les pratiques de la taxonomie qui a causé un certain refus de ces technologies par les taxonomistes. Les vives critiques dont ils ont fait l'objet les ont menés à argumenter sur l'importance de leur discipline et son historique, menant ainsi à un « dialogue de sourds » entre les taxonomistes et les autres biologistes, pour reprendre l'expression d'Anouk Barberousse et Sarah Samadi (Barberousse & Samadi, 2013). L'arrivée du séquençage ADN, en particulier, a été l'objet de vifs débats dans la discipline ; les convaincus de ces méthodes ont ainsi dû développer des stratégies de persuasion pour obtenir des morphologistes qu'ils changent leurs techniques de travail et de conservation des spécimens, afin de rendre possible les analyses ADN (Mauz & Faugère, 2013).

La synthèse entre méthode morphologique et moléculaire permise par la taxonomie intégrative fait toutefois la part belle aux outils et données numériques : si la comparaison des données morphologiques peut être automatisée sur de grands jeux de données, qui ne sont toutefois pas encore la norme, les données issues des séquençages ADN sont numériques par nature et peuvent être facilement comparés en utilisant les outils adéquats. Notons aussi que le spécimen numérique étendu, même si la notion est récente, implique une utilisation poussée des outils numériques, avec des infrastructures techniques qui nécessitent une forte organisation des données, aussi bien sémantique que technique : le TDWG a ainsi consacré un séminaire entier à la question en 2021, ainsi

que des séminaires sur la question des données de la biodiversité depuis plusieurs années.

Il faut cependant remarquer que les bases de données et les ressources spécialisées sont disponibles sur Internet, qui est devenu la meilleure façon de partager les informations scientifiques entre chercheurs, conformément à sa vocation initiale. Les mouvements de l'open science et de l'open data ont ainsi fait émerger les problématiques d'accès aux produits et aux données de la recherche sur le court et moyen terme. Ainsi, la mise en ligne des données de la recherche pourrait être une porte d'entrée pour amener la question de l'archivage pérenne. Le principe des *data papers*, dont l'enjeu est de rendre les données aussi importantes qu'une autre publication dans la carrière des scientifiques, pourrait ainsi pousser à une meilleure qualité des données ; leur description précise peut également correspondre, dans une certaine mesure, aux descriptions archivistiques car ils rappellent de façon extensive les traitements effectués sur les données ainsi que leur contexte de production.

3) Des débuts encourageants

Il nous semble toutefois que malgré les obstacles à l'archivage des données de la biodiversité, des efforts sont faits pour inscrire les bases de données de la biodiversité dans les standards techniques.

L'une des pistes concerne le Web Sémantique et son application au domaine de la biodiversité : la plateforme Plazi[28] propose ainsi de relier les traitements taxonomiques présents dans les articles qu'elle indexe selon l'ontologie OpenBioDiv-O. Cette dernière a également été utilisée pour produire un graphe de connaissances reprenant des informations présentes sur Plazi, sur le GBIF et d'autres publiées sur Pensoft (Penev et al., 2019) : il s'agit alors de moissonner la littérature scientifique afin de permettre une exploration automatisée des taxons publiés. Notons aussi que l'ABCD est disponible, depuis sa version 3.0, en tant qu'ontologie[29].

La question des identifiants est un des enjeux de la mise en ligne des données, et elle se pose peut-être de façon encore plus accrue lorsqu'il s'agit du web sémantique. Ainsi, le CETAF (Consortium of European Taxonomic Facilites) a décidé, depuis 2013, d'utiliser les URI au format HTTP pour l'attribution d'identifiants aux ressources mises en lignes par les différentes institutions membres du consortium (Güntsch et al., 2017). L'un des exemples d'utilisation concerne les récolteurs de plantes conservés dans les herbiers, qui ont été annotés sémantiquement et permet de regrouper des spécimens qui ont été collectés par les mêmes personnes (Güntsch et al., 2021).

Il nous semble toutefois difficile d'appliquer le même principe aux noms scientifiques, même si ceux-ci jouent globalement le rôle d'identifiants uniques pour un même taxon au sein de la communauté scientifique : parce qu'ils représentent des hypothèses taxonomiques, ils sont sujets à révision et ne peuvent donc être considérés comme des identifiants pérennes. Appliquer un identifiant unique à un spécimen naturalisé semble toutefois à portée de main, dans la mesure

[28] *Plazi*. (s. d.). Consulté 22 août 2023, à l'adresse https://plazi.org/

[29] *Project, A. 3 0.* (s. d.). Abcd—Ontology primer. Consulté 22 août 2023, à l'adresse https://abcd.tdwg.org/ontology/documentation/primer /

où il s'agit de pouvoir faire référence à un individu précis et unique, qui est conservé sur le plus ou moins long terme, et non d'attribuer un identifiant pérenne à un concept susceptible d'évoluer avec le temps.

Nous aimerions toutefois noter que ces pratiques, si elles sont utilisées dans l'archivage, n'en font pas partie à proprement parler ; on pourrait même penser que cela complique le travail de l'archiviste, dans le sens où la description en RDF ne ferait que redonder des informations déjà présente ailleurs. À notre avis, l'appropriation des technologies du web sémantique par la biologie peut être considérée comme un point d'entrée pour sensibiliser les chercheurs à l'importance de l'archivage de leurs données, en plus d'inscrire les données de la biodiversité dans un formalisme de représentation des connaissances émergent qu'est le Web Sémantique : ce dernier a pour but de relier entre elles les données présentes sur internet, de façon à ce qu'elles soient à la fois lisibles pour les humains et exploitables automatiquement par les machines, tout en rajoutant un niveau de signification sous forme de triplets RDF. Des ontologies existent pour les données de la biodiversité : nous avons déjà évoqué l'ABCD, mais le Darwin Core est lui aussi développé sous forme d'un graphe RDF (Wieczorek et al., 2012), en plus de sa version XML qui reste la plus utilisée.

Certains articles posent toutefois la question des financements sur le long terme des données de la recherche en biologie, qui sont essentielles au maintien en ligne d'une base de données, peu importe la spécificité des données qui y sont déposées : une étude montre ainsi la répartition des financements entre acteurs gouvernementaux, académiques et philanthropiques pour les bases de données moléculaires (Imker, 2020).

Il nous semble ainsi qu'il y a peu d'institutions qui portent un intérêt à la conservation sur le long terme des données de la biodiversité : pour l'instant les efforts nous semblent tournés, à l'instar des objectifs du projet européen BICIKL (BIodiversity Community Integrated Knowledge Library)[30], sur la mise à disposition de données de qualité et non sur leur archivage ; ce portail a pour vocation d'agir comme une porte d'entrée vers différents services proposés par les infrastructures de recherche et partenaires de recherche membres. La démarche relève davantage, selon nous, de celle portée par les bibliothèques que les archives, peut-être parce que les données de la biodiversité ne sont pas considérées par les biologistes comme des archives potentielles.

[30] *Bicikl Homepage*. (s. d.). Consulté 25 août 2023, à l'adresse https://bicikl-project.eu/

CONCLUSION

Prendre la taxonomie comme fil rouge de ce mémoire nous a conduit à aborder différents points. Tout d'abord, nous avons vu que les pratiques scientifiques ont changé, et qu'à cet égard la fin des années 1990 et le début des années 2000 ont été cruciales. Le « *taxonomic impediment* », cet handicap taxonomique, a forcé les taxonomistes à trouver de nouvelles façons de mener leur recherche scientifique, et à réaffirmer leur place au sein de la biologie ; ce renouvellement de la taxonomie est passé par la prise en main de nouveaux outils et de nouvelles pratiques. L'arrivée du séquençage ADN a joué le rôle d'un catalyseur, à la fois des critiques à l'encontre de la taxonomie et des solutions, même si ces dernières ne pouvaient pas raisonnablement « faire sans » la méthode qui l'a précédé. Le passage de la taxonomie morphologique à la taxonomie moléculaire ne se serait pas fait sans la fusion permise par la taxonomie intégrative, et celle-ci a également du mal à s'imposer dans les pratiques, même si les acteurs institutionnels et scientifiques peuvent parfois pousser à l'adoption d'outils, selon leurs propres agendas. L'invariant reste toutefois l'importance accordée au spécimen physique, seule preuve définitive d'une observation (mais pas d'une identification scientifique). Toutefois, les pratiques qui mènent au spécimen ont changé : le séquençage ADN nécessite un conditionnement de l'individu bien particulier pour préserver les tissus où l'extraction du *barcode* est à la fois le plus aisé et le plus fiable. La photographie des individus, en remplacement du prélèvement de l'un d'entre eux, change également la donne, puisqu'il n'y a plus moyen de vérifier précisément l'identification scientifique d'un animal au moyen de sa morphologie.

Ces nouvelles pratiques ne sont pas sans changer les fondements mêmes de la taxonomie : pour s'emparer totalement les outils qu'on leur présentait, les taxonomistes ont fini par élargir leur objet de recherche initial. En effet, le spécimen, d'abord objet unique dans des collections naturalistes, s'est vu tour à tour étendu, numérisé, numérique, pour finir étendu et numérique. Son spectre s'est élargi, d'autres données, d'autres documents se joignent à lui pour former un nouvel ensemble qui ouvre d'autres possibilités de recherche. L'objet physique matérialisé sous une autre forme nous semble alors changer en partie de régime : d'abord exemplaire unique, les données qui le composent peuvent être facilement recopiées et diffusées, au prix d'une perte dans la mesure où de nos jours les réalisations qui se rapprochent le plus du spécimen numérique étendu ne dispensent pas les chercheurs de consulter l'objet en vrai. Il nous a alors semblé intéressant d'étudier dans quelle mesure la relation au spécimen physique a changé avec le développement du spécimen numérique étendu, de la même façon que le rapport aux archives a pu être modifié avec la mise en ligne massive des documents anciens.

La fonction documentaire du spécimen, numérique ou non, étendu ou non, est ce qui nous a mené à étudier sa conservation sur le long terme. Document unique par excellence (puisqu'il s'agit après tout de conserver un individu) , le spécimen une fois collecté nous renseigne, pour peu qu'il soit suffisamment documenté, sur le monde qui l'entoure : témoin, trace, donnée, d'une certaine manière, dans un (éco)système que les biologistes cherchent à étudier dans les moindre détail. La question de l'archivage, ou du moins de la conservation sur le long terme de ces données, n'est toutefois pas considérée en tant que telle par les

chercheurs. De son côté, le monde de la documentation a laissé aux biologistes le soin de s'occuper des spécimens ; les désintérêts sont présents de part et d'autre. Pourtant, des ponts existent pour relier les deux mondes, dans la mesure où la diffusion des données de la recherche est une question qui concerne à la fois les scientifiques (en tant que producteurs et potentiels réutilisateurs des données), les bibliothécaires (sur le volet de l'accès aux données) et les archivistes (assurer la transmission des données intactes, que ce soit sur un court laps de temps ou une durée beaucoup plus longue). Faire des liens entre ces trois communautés est un défi qu'il nous semble nécessaire de relever pour assurer la bonne transmission des données de la recherche que constitue le spécimen numérique étendu.

Enfin, et nous voudrions terminer là-dessus, il nous semble important d'insister une dernière fois sur les enjeux de la transmission de nos connaissances actuelles sur le vivant qui nous entoure. « Sans mémoire, pas d'histoire », écrit Bruno Bachimont en page 11 de son ouvrage Patrimoine et numérique[31]. Patrimonialiser les objets qui permettent de faire de l'histoire naturelle, avec ce que cela implique de concepts et de pratiques, permettrait de rendre ces objets disponibles sur le très long terme ; rendre ces objets disponibles dans l'univers numérique, avec ce que cela suppose de médiations techniques, c'est toutefois risquer de ne plus pouvoir y accéder. « Sans mémoire du vivant, pas d'histoire naturelle » : voilà comment on pourrait adapter la citation de Bruno Bachimont à notre sujet.

[31]Bachimont, B. (2017). *Patrimoine et numérique : Technique et politique de la mémoire.* INA.

SOURCES

Entretiens par téléphone et en visioconférence :

- M. Aurélien Mirallès, taxonomiste à l'Institut de Systématique, Évolution, Biodiversité (UMR 7205 – CNRS – MNHN – EPHE-PSL – Université des Antilles), le 24 avril 2023.
- Mme Roselli Pellens, macroécologiste à l'Institut de Systématique, Évolution, Biodiversité (UMR 7205 – CNRS – MNHN – EPHE-PSL – Université des Antilles), le 20 juin 2023.
- Mme Sophie Pamerlon, ingénieure données au GBIF, le 27 juin 2023.

Pages web consultées, dans l'ordre d'apparition

Access to biological collection data (Abcd) schema. (s. d.). Consulté 18 août 2023, à l'adresse https://www.tdwg.org/standards/abcd/

Biodiversity information standards(Tdwg). (s. d.). Consulté 18 août 2023, à l'adresse https://www.tdwg.org/

TDWG: History. (s. d.). TDWG. Consulté 18 août 2023, à l'adresse https://web.archive.org/web/20180329125359/http://www.tdwg.org/about-tdwg/history/

Analytics, C. (s. d.). ION : Index to Organism Names [Document]. Consulté 26 août 2023, à l'adresse http://www.organismnames.com/metrics.htm?page=tsj

Museum Koenig Bonn. (s. d.). Sammlungen | Museum Koenig Bonn. Forschungsmuseum Koenig ; Museum Koenig Bonn. Consulté 13 août 2023, à l'adresse https://bonn.leibniz-lib.de/de/forschung/sammlungen

Museum Koenig Bonn. (s. d.). Digitalisierungsstrategie für die wissenschaftlichen Sammlungen am ZFMK. Forschungsmuseum Koenig ; Museum Koenig Bonn.

Digital collection catalogue statistics. (s. d.). Consulté 13 août 2023, à l'adresse https://collections.leibniz-lib.de/statistics/

Natural Sciences collection areas. (s. d.). The Australian Museum. Consulté 13 août 2023, à l'adresse https://australian.museum/learn/collections/natural-science/australian.museum/learn/collections/natural-science/

Welcome—Data portal. (s. d.). Consulté 13 août 2023, à l'adresse https://data.nhm.ac.uk/

SI NMNH—Museum collection search. (s. d.). Consulté 13 août 2023, à l'adresse https://collections.nmnh.si.edu/search

Museum collections policies | Smithsonian National Museum of Natural History. (s. d.). Consulté 13 août 2023, à l'adresse http://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

Qu'est-ce que le Muséum ? (s. d.). Muséum national d'Histoire naturelle. Consulté 13 août 2023, à l'adresse https://www.mnhn.fr/fr/qu-est-ce-que-le-museum

TDWG 2021 : SYM07 Digital Extended Specimens. (2023). Consulté 25 août 2023, à l'adresse https://www.youtube.com/watch?v=jpY7vh7BgKM à 01:13:53

La Planète Revisitée. (s. d.). Muséum national d'Histoire naturelle. Consulté 16 août 2023, à l'adresse https://www.mnhn.fr/fr/la-planete-revisitee

Biodiversity Heritage Library. (s. d.). Consulté 24 août 2023, à l'adresse https://www.biodiversitylibrary.org/

Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., Stjernegaard Jeppesen, T., Örn, A., Vandepitte, L., Hobern, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Abbott, J., Adlard, R., Adriaenssens, E. M., et al. (2023). *Catalogue of Life Checklist* (Version 2023-07-18) [Dataset]. Catalogue of Life. https://doi.org/10.48580/dfsy

Pomart, J. (2014, juillet 18). AAF / Section Aurore : Un groupe de travail sur les données de la recherche [Billet]. *Archives de la FMSH.* https://archivesfmsh.hypotheses.org/1209

Genbank overview. (s. d.). Consulté 20 août 2023, à l'adresse https://www.ncbi.nlm.nih.gov/genbank/

Ena browser. (s. d.). Consulté 20 août 2023, à l'adresse https://www.ebi.ac.uk/ena/browser/home

Col. (s. d.). COL. Consulté 20 août 2023, à l'adresse https://www.catalogueoflife.org/

Source datasets. (s. d.). COL. Consulté 20 août 2023, à l'adresse https://www.catalogueoflife.org/data/source-datasets

Gbif. (s. d.). Consulté 23 août 2023, à l'adresse https://www.gbif.org/

Dryad | Home—Publish and preserve your data. (s. d.). Dryad. Consulté 20 août 2023, à l'adresse https://datadryad.org/stash

Plazi. (s. d.). Consulté 22 août 2023, à l'adresse https://plazi.org/

Project, A. 3 0. (s. d.). Abcd—Ontology primer. Consulté 22 août 2023, à l'adresse https://abcd.tdwg.org/ontology/documentation/primer /

Bicikl Homepage. (s. d.). Consulté 25 août 2023, à l'adresse https://bicikl-project.eu/

BIBLIOGRAPHIE

Introduction et conclusion

Bachimont, B. (2017). *Patrimoine et numérique : Technique et politique de la mémoire.* INA.

Cowie, R. H., Bouchet, P., & Fontaine, B. (2022). The Sixth Mass Extinction : Fact, fiction or speculation? *Biological Reviews*, *97*(2), 640-663. https://doi.org/10.1111/brv.12816

Gadelha, L. M. R., Siracusa, P. C., Dalcin, E. C., Silva, L. A. E., Augusto, D. A., Krempser, E., Affe, H. M., Costa, R. L., Mondelli, M. L., Meirelles, P. M., Thompson, F., Chame, M., Ziviani, A., & Siqueira, M. F. (2021). A survey of biodiversity informatics : Concepts, practices, and challenges. *WIREs Data Mining and Knowledge Discovery*, *11*(1). https://doi.org/10.1002/widm.1394

I. La taxonomie, histoire et contexte

Agnarsson, I., & Kuntner, M. (2007). Taxonomy in a Changing World : Seeking Solutions for a Science in Crisis. *Systematic Biology*, *56*(3), 531-539. https://doi.org/10.1080/10635150701424546

Barabé, D., Cuerrier, A., & Quilichini, A. (2012). Les jardins botaniques : Entre science et commercialisation. *Natures Sciences Sociétés*, *20*(3), 334-342. https://doi.org/10.1051/nss/2012040

Barberousse, A., & Samadi, S. (2013). La taxonomie dans la tourmente. *Revue d'anthropologie des connaissances*, *7*(2). https://doi.org/10.3917/rac.019.0411

Browne, J. (1997). Une science impérialisme : L'histoire naturelle britannique et les voyages d'exploration de Banks à Darwin. In C. Blanckaert, C. Cohen, P. Corsi, & J.-L. Fischer (Éds.), *Le Muséum au premier siècle de son histoire* (p. 197-210). Publications scientifiques du Muséum. https://doi.org/10.4000/books.mnhn.1699

Costa, F. O., & Carvalho, G. R. (2007). The Barcode of Life Initiative : Synopsis and prospective societal impacts of DNA barcoding of Fish. *Genomics, Society and Policy*, *3*(2), 29. https://doi.org/10.1186/1746-5354-3-2-29

Dayrat, B. (2005). Towards integrative taxonomy : INTEGRATIVE TAXONOMY. *Biological Journal of the Linnean Society*, *85*(3), 407-415. https://doi.org/10.1111/j.1095-8312.2005.00503.x

Draelants, I. (1996). *Les encyclopédies comme sommes des connaissances, d'Isidore de Séville au XIIIe siècle* (Vol. 2, p. 25). https://shs.hal.science/halshs-03095401

Ellis, R., Waterton, C., & Wynne, B. (2010). Taxonomy, biodiversity and their publics in twenty-first-century DNA barcoding. *Public Understanding of Science*, *19*(4), 497-512. https://doi.org/10.1177/0963662509335413

Funk, V. A. (2001). SSZ 1970–1989 : A View of the Years of Conflict. *Systematic Biology*, *50*(2), 153-155. https://doi.org/10.1080/10635150151125798

Gemeinholzer, B., Vences, M., Beszteri, B., Bruy, T., Felden, J., Kostadinov, I., Miralles, A., Nattkemper, T. W., Printzen, C., Renz, J., Rybalka, N., Schuster, T., Weibulat, T., Wilke, T., & Renner, S. S. (2020). Data storage and data re-use in taxonomy—The need for improved storage and accessibility of heterogeneous data. *Organisms Diversity & Evolution*, *20*(1), 1-8. https://doi.org/10.1007/s13127-019-00428-w

Godfray, H. C. J., & Knapp, S. (2004). Introduction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 559-569. https://doi.org/10.1098/rstb.2003.1457

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313-321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of Birds through DNA Barcodes. *PLoS Biology*, *2*(10), e312. https://doi.org/10.1371/journal.pbio.0020312

International Commission on Zoological Nomenclature, Ride, W. D. L., International Trust for Zoological Nomenclature, Natural History Museum (London, England), & International Union of Biological Sciences (Éds.). (1999). *International code of zoological nomenclature = : Code international de nomenclature zoologique* (4th ed). International Trust for Zoological Nomenclature, c/o Natural History Museum. https://www.iczn.org/the-code/the-code-online/

Jozwiak, P., Rewicz, T., & Pabis, K. (2015). Taxonomic etymology – in search of inspiration. *ZooKeys*, *513*, 143-160. https://doi.org/10.3897/zookeys.513.9873

Kemp, C. (2015). Museums : The endangered dead. *Nature*, *518*(7539), 292-294. https://doi.org/10.1038/518292a

Larson, B. M. (2007). DNA Barcoding : The Social Frontier. *Frontiers in Ecology and the Environment*, *5*(8), 437-442.

Lipscomb, D., Platnick, N., & Wheeler, Q. (2003). The intellectual content of taxonomy : A comment on DNA taxonomy. *Trends in Ecology & Evolution*, *18*(2), 65-66. https://doi.org/10.1016/S0169-5347(02)00060-5

Mauz, I., & Faugère, E. (2013). Les systématiciens à l'épreuve du barcoding : Une étude des pratiques d'enrôlement scientifique. *Revue d'anthropologie des connaissances*, *7*(2). https://doi.org/10.3917/rac.019.0433

Miller, S. E. (2007). DNA barcoding and the renaissance of taxonomy. *Proceedings of the National Academy of Sciences*, *104*(12), 4775-4776. https://doi.org/10.1073/pnas.0700466104

Moritz, C., & Cicero, C. (2004). DNA Barcoding : Promise and Pitfalls. *PLoS Biology*, *2*(10), e354. https://doi.org/10.1371/journal.pbio.0020354

Padial, J. M., Miralles, A., De La Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, *7*(1), 16. https://doi.org/10.1186/1742-9994-7-16

Parra-Olea, G., Rovito, S. M., García-París, M., Maisano, J. A., Wake, D. B., & Hanken, J. (2016). Biology of tiny animals : Three new species of minute salamanders (Plethodontidae: *Thorius*) from Oaxaca, Mexico. *PeerJ*, *4*, e2694. https://doi.org/10.7717/peerj.2694

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends in Ecology & Evolution*, *36*(11), 1049-1060. https://doi.org/10.1016/j.tree.2021.07.009

Robert, P. (2010). *Mnémotechnologies : Une théorie générale critique des technologies intellectuelles*. Hermès science publications : Lavoisier.

Sauvage, Ch. (1965). Taxinomie ou taxonomie. *Bulletin de la Société Botanique de France*, *112*(3-4), 180-182. https://doi.org/10.1080/00378941.1965.10838227

Sennikov, A. N., & Lazkov, G. A. (2023). Taxonomic revision of the *Allium filidens* group (Amaryllidaceae) in Kyrgyzstan. *Nordic Journal of Botany*, e04050. https://doi.org/10.1111/njb.04050

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2002). DNA points the way ahead in taxonomy. *Nature*, *418*(6897), 479-479. https://doi.org/10.1038/418479a

Tillier, S. (2005). Terminologie et nomenclatures scientifiques : L'exemple de la taxonomie zoologique: *Langages*, *n° 157*(1), 104-117. https://doi.org/10.3917/lang.157.0104

Turland, N., Wiersema, J., Barrie, F., Greuter, W., Hawksworth, D., Herendeen, P., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T., McNeill, J., Monro, A., Prado, J., Price, M., & Smith, G. (Éds.). (2018). *International Code of Nomenclature for algae, fungi, and plants* (Vol. 159). Koeltz Botanical Books. https://doi.org/10.12705/Code.2018

Vences, M., Miralles, A., Brouillet, S., Ducasse, J., Fedosov, A., Kharchev, V., Kostadinov, I., Kumari, S., Patmanidis, S., Scherz, M. D., Puillandre, N., & Renner, S. S. (2021). iTaxoTools 0.1 : Kickstarting a specimen-based software toolkit for taxonomists. *Megataxa*, *6*(2). https://doi.org/10.11646/megataxa.6.2.1

Whitfield, J. (2003). DNA barcodes catalogue animals. *Nature*, news030512-7. https://doi.org/10.1038/news030512-7

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core : An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

II. Le *Digital Extended Specimen*

Briet, S. (1951). *Qu'est-ce que la documentation ?* EDIT.

Deloitte Economics. (2023). *Economic valuation and assessment of the impact of the GBIF network*. https://www.deloitte.com/content/dam/assets-zone1/au/en/docs/services/economics/deloitte-economics-global-biodiversity-information-facility-260623.pdf

Groom, Q., Hyam, R., & Güntsch, A. (2017). Stable identifiers for collection specimens. *Nature*, *546*(7656), 33-33. https://doi.org/10.1038/546033d

Hansen, S., Franz, N., & Monfils, A. (2022). Early Career Scientists are Critical to the FAIR Data Pathway. *Biodiversity Information Science and Standards*, *6*, e90989. https://doi.org/10.3897/biss.6.90989

Hardisty, A. R., Ellwood, E. R., Nelson, G., Zimkus, B., Buschbom, J., Addink, W., Rabeler, R. K., Bates, J., Bentley, A., Fortes, J. A. B., Hansen, S., Macklin, J. A., Mast, A. R., Miller, J. T., Monfils, A. K., Paul, D. L., Wallis, E., & Webster, M. (2022). Digital Extended Specimens : Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. *BioScience*, *72*(10), 978-987. https://doi.org/10.1093/biosci/biac060

Hardisty, A. R., Saarenmaa, H., Casino, A., Dillen, M., Gödderz, K., Groom, Q., Hardy, H., Koureas, D., Hidalga, A. N. de la, Paul, D. L., Runnel, V., Vermeersch, X., Walsum, M. van, & Willemse, L. (2020). Conceptual design blueprint for the DiSSCo digitization infrastructure—DELIVERABLE D8.1. *Research Ideas and Outcomes*, *6*, e54280. https://doi.org/10.3897/rio.6.e54280

Kays, R., McShea, W. J., & Wikelski, M. (2020). Born-digital biodiversity data : Millions and billions. *Diversity and Distributions*, *26*(5), 644-648. https://doi.org/10.1111/ddi.12993

Lendemer, J., Thiers, B., Monfils, A. K., Zaspel, J., Ellwood, E. R., Bentley, A., LeVan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L., Mabee, P., Ford, L. S., Guralnick, R., Gropp, R. E., Revelez, M., Cobb, N., Seltmann, K., & Aime, M. C. (2020). The Extended Specimen Network : A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. *BioScience*, *70*(1), 23-30. https://doi.org/10.1093/biosci/biz140

Leonelli, S. (2010). Packaging Small Facts for Re-Use : Databases in Model Organism Biology: In P. Howlett & M. S. Morgan (Éds.), *How Well Do Facts Travel?* (1re éd., p. 325-348). Cambridge University Press. https://doi.org/10.1017/CBO9780511762154.017

Leonelli, S., & Ankeny, R. A. (2012). Re-thinking organisms : The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 29-36. https://doi.org/10.1016/j.shpsc.2011.10.003

Monfils, A., & Ellwood, E. R. (2021). Workforce Capacity Development and the Digital Extended Specimen. *Biodiversity Information Science and Standards*, *5*, e73927. https://doi.org/10.3897/biss.5.73927

Moretti, F. (2000). Conjectures on World Literature. *New Left Revew*, *1*(1). https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Nelson, G., Paul, D., Riccardi, G., & Mast, A. (2012). Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys*, *209*, 19-45. https://doi.org/10.3897/zookeys.209.3135

Tegelberg, R., Mononen, T., & Saarenmaa, H. (2014). High-performance digitization of natural history collections : Automated imaging lines for herbarium and insect specimens. *TAXON*, *63*(6), 1307-1313. https://doi.org/10.12705/636.13

Uhen, M. D., Barnosky, A. D., Bills, B., Blois, J., Carrano, M. T., Carrasco, M. A., Erickson, G. M., Eronen, J. T., Fortelius, M., Graham, R. W., Grimm, E. C., O'Leary, M. A., Mast, A., Piel, W. H., Polly, P. D., & Säilä, L. K. (2013). From card catalogs to computers : Databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, *33*(1), 13-28. https://doi.org/10.1080/02724634.2012.716114

Webster, M., Buschbom, J., Hardisty, A., & Bentley, A. (2021). The Digital Extended Specimen will Enable New Science and Applications. *Biodiversity Information Science and Standards*, *5*, e75736. https://doi.org/10.3897/biss.5.75736

Webster, M. S. (Éd.). (2018). *The extended specimen : Emerging frontiers in collections-based ornithological research*. CRC Press, Taylor & Francis Group.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zamani, A., Fric, Z. F., Gante, H. F., Hopkins, T., Orfinger, A. B., Scherz, M. D., Bartoňová, A. S., & Pos, D. D. (2022). DNA barcodes on their own are not enough to describe a species. *Systematic Entomology*, *47*(3), 385-389. https://doi.org/10.1111/syen.12538

III. L'archivage des données de la biodiversité, entre données numériques et spécimens physiques

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment : Working with scientists to understand data management and preservation needs. *Archival Science*, *11*(3-4), 329-348. https://doi.org/10.1007/s10502-011-9151-4

Barberousse, A., & Samadi, S. (2013). La taxonomie dans la tourmente. *Revue d'anthropologie des connaissances*, *7*(2). https://doi.org/10.3917/rac.019.0411

Charmasson, T. (2006). Archives scientifiques ou archives des sciences : Des sources pour l'histoire. *La Revue pour l'histoire du CNRS*, *14*. https://doi.org/10.4000/histoire-cnrs.1790

Colella, J. P., Stephens, R. B., Campbell, M. L., Kohli, B. A., Parsons, D. J., & Mclean, B. S. (2021). The Open-Specimen Movement. *BioScience*, *71*(4), 405-414. https://doi.org/10.1093/biosci/biaa146

Farnham, A., Kurz, C., Öztürk, M. A., Solbiati, M., Myllyntaus, O., Meekes, J., Pham, T. M., Paz, C., Langiewicz, M., Andrews, S., Kanninen, L., Agbemabiese, C., Guler, A. T., Durieux, J., Jasim, S., Viessmann, O., Frattini, S., Yembergenova, D., Benito, C. M., … Hettne, K. (2017). Early career researchers want Open Science. *Genome Biology*, *18*(1), 221. https://doi.org/10.1186/s13059-017-1351-7

Gadelha, L. M. R., Siracusa, P. C., Dalcin, E. C., Silva, L. A. E., Augusto, D. A., Krempser, E., Affe, H. M., Costa, R. L., Mondelli, M. L., Meirelles, P. M., Thompson, F., Chame, M., Ziviani, A., & Siqueira, M. F. (2021). A survey of biodiversity informatics : Concepts, practices, and challenges. *WIREs Data Mining and Knowledge Discovery*, *11*(1). https://doi.org/10.1002/widm.1394

Groom, Q., Adriaens, T., Bertolino, S., Phelps, K., Poelen, J., Reeder, D., Richardson, D., Simmons, N., Trekels, M., & Upham, N. (2022). The Importance of Collecting and Archiving Data on Domestic and Cultivated Organisms. *Biodiversity Information Science and Standards*, *6*, e90864. https://doi.org/10.3897/biss.6.90864

Guimont, C., & Petitimbert, R. (2017). Instruments de l'action publique et approche fixiste de la biodiversité : Le cas des inventaires naturalistes. *Norois. Environnement, aménagement, société*, *244*, Article 244. https://doi.org/10.4000/norois.6169

Güntsch, A., Groom, Q., Ernst, M., Holetschek, J., Plank, A., Röpert, D., Fichtmüller, D., Shorthouse, D. P., Hyam, R., Dillen, M., Trekels, M., Haston, E., & Rainer, H. (2021). A botanical demonstration of the potential of linking data using unique identifiers for people. *PLOS ONE*, *16*(12), e0261130. https://doi.org/10.1371/journal.pone.0261130

Güntsch, A., Groom, Q., Hyam, R., Chagnoux, S., Röpert, D., Berendsohn, W. G., Casino, A., Droege, G., Gerritsen, W., Holetschek, J., Marhold, K., Mergen, P., Rainer, H., Smith, V., & Triebel, D. (2018). Standardised Globally Unique

Specimen Identifiers. *Biodiversity Information Science and Standards*, *2*, e26658. https://doi.org/10.3897/biss.2.26658

Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., Droege, G., Glöckler, F., Gödderz, K., Groom, Q., Hoffmann, J., Holleman, A., Kempa, M., Koivula, H., Marhold, K., Nicolson, N., Smith, V. S., & Triebel, D. (2017). Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*, *2017*. https://doi.org/10.1093/database/bax003

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., & Holleley, C. E. (2022). Unlocking inaccessible historical genomes preserved in formalin. *Molecular Ecology Resources*, *22*(6), 2130-2147. https://doi.org/10.1111/1755-0998.13505

Imker, H. J. (2020). Who Bears the Burden of Long-Lived Molecular Biology Databases? *Data Science Journal*, *19*(1), 8. https://doi.org/10.5334/dsj-2020-008

Johnson, K. R., Owens, I. F. P., & the Global Collection Group. (2023). A global approach for natural history museum collections. *Science*, *379*(6638), 1192-1194. https://doi.org/10.1126/science.adf6434

Larsen, B. B., Miller, E. C., Rhodes, M. K., & Wiens, J. J. (2017). Inordinate Fondness Multiplied and Redistributed : The Number of Species on Earth and the New Pie of Life. *The Quarterly Review of Biology*, *92*(3), 229-265. https://doi.org/10.1086/693564

Lister, A. M. (2011). Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution*, *26*(4), 153-154. https://doi.org/10.1016/j.tree.2010.12.009

Marshall, C. R., Finnegan, S., Clites, E. C., Holroyd, P. A., Bonuso, N., Cortez, C., Davis, E., Dietl, G. P., Druckenmiller, P. S., Eng, R. C., Garcia, C., Estes-Smargiassi, K., Hendy, A., Hollis, K. A., Little, H., Nesbitt, E. A., Roopnarine, P., Skibinski, L., Vendetti, J., & White, L. D. (2018). Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biology Letters*, *14*(9), 20180431. https://doi.org/10.1098/rsbl.2018.0431

Mauz, I., & Faugère, E. (2013). Les systématiciens à l'épreuve du barcoding : Une étude des pratiques d'enrôlement scientifique. *Revue d'anthropologie des connaissances*, *7*(2). https://doi.org/10.3917/rac.019.0433

Mesguich, V. (2023). Chapitre 5. Les données de la recherche. In *Les bibliothèques face au monde des données* (p. 91-110). Presses de l'enssib. https://doi.org/10.4000/books.pressesenssib.17768

Minteer, B. A., Collins, J. P., & Puschendorf, R. (2014). Specimen collection : An essential tool—Response. *Science*, *344*(6186), 816-816. https://doi.org/10.1126/science.344.6186.816-a

Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., Beszteri, B., Bonkowski, M., Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F. O., Hawlitschek, O., Kostadinov, I., Nattkemper, T. W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., … Vences, M. (2020). Repositories for Taxonomic Data : Where We Are and What is Missing. *Systematic Biology*, *69*(6), 1231-1253. https://doi.org/10.1093/sysbio/syaa026

Mora-Cross, M., Morales-Carmiol, A., Chen-Huang, T., & Barquero-Pérez, M. (2022). Essential Biodiversity Variables : Extracting plant phenological data from specimen labels using machine learning. *Research Ideas and Outcomes*, *8*, e86012. https://doi.org/10.3897/rio.8.e86012

Mulcahy, D. G., Ibáñez, R., Jaramillo, C. A., Crawford, A. J., Ray, J. M., Gotte, S. W., Jacobs, J. F., Wynn, A. H., Gonzalez-Porter, G. P., McDiarmid, R. W., Crombie, R. I., Zug, G. R., & de Queiroz, K. (2022). DNA barcoding of the National Museum of Natural History reptile tissue holdings raises concerns about the use of natural history collections and the responsibilities of scientists in the molecular age. *PLOS ONE*, *17*(3), e0264930. https://doi.org/10.1371/journal.pone.0264930

Penev, L., Dimitrova, M., Senderov, V., Zhelezov, G., Georgiev, T., Stoev, P., & Simov, K. (2019). OpenBiodiv : A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. *Publications*, *7*(2), 38. https://doi.org/10.3390/publications7020038

Pernet, C., Svarer, C., Blair, R., Van Horn, J. D., & Poldrack, R. A. (2023). On the Long-term Archiving of Research Data. *Neuroinformatics*, *21*(2), 243-246. https://doi.org/10.1007/s12021-023-09621-x

Poisot, T., Mounce, R., & Gravel, D. (2013). Moving toward a sustainable ecological science : Don't let data go to waste! *Ideas in Ecology and Evolution*, *6*(2). https://doi.org/10.4033/iee.2013.6b.14.f

Pomart, J. (2014, juillet 18). AAF / Section Aurore : Un groupe de travail sur les données de la recherche [Billet]. *Archives de la FMSH*. https://archivesfmsh.hypotheses.org/1209

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends in Ecology & Evolution*, *36*(11), 1049-1060. https://doi.org/10.1016/j.tree.2021.07.009

Robert, P. (2010). *Mnémotechnologies : Une théorie générale critique des technologies intellectuelles*. Hermès science publications : Lavoisier.

Rocha, L. A., Aleixo, A., Allen, G., Almeda, F., Baldwin, C. C., Barclay, M. V. L., Bates, J. M., Bauer, A. M., Benzoni, F., Berns, C. M., Berumen, M. L., Blackburn, D. C., Blum, S., Bolaños, F., Bowie, R. C. K., Britz, R., Brown, R. M., Cadena, C.

D., Carpenter, K., … Witt, C. C. (2014). Specimen collection : An essential tool. *Science*, *344*(6186), 814-815. https://doi.org/10.1126/science.344.6186.814

Staunton, C., Barragán, C. A., Canali, S., Ho, C., Leonelli, S., Mayernik, M., Prainsack, B., & Wonkham, A. (2021). Open science, data sharing and solidarity : Who benefits? *History and Philosophy of the Life Sciences*, *43*(4), 115. https://doi.org/10.1007/s40656-021-00468-6

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. https://doi.org/10.1093/sysbio/syy044

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core : An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, *7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

ANNEXES

Table des annexes

Traduction de l'illustration n°1

Allium filidentiforme Vved. (Vvedensky, 1925, p. 32)

Type : Kirghizstan, Région nord des Monts Suzak, 4 juillet 1945, Kalinina and Moreva 71 (holotype: LE).

Herbe vivace à bulbe, sans rhizomes. Bulbes ovoïdes, jusqu'à 2,5 cm de diamètre. Tunique extérieure brune, réticulée ; haut du bulbe ne dépassant pas du sol ; tunique intérieure fine comme du papier avec des veines visibles. Bulbilles rares et solitaires. Tiges hautes de 50-70 cm, 5 mn de diamètre, térétiformes, lisses, couvertes de gaines foliaires sur 1/3 de la longueur. 3-4 feuilles, ressemblant aux roseaux, espacées, brins longs de 20-40 cm, 3 mn de diamètre, s'effilant jusqu'à la pointe, lisses, verts. Inflorescence de 5 cm de diamètre, globulaire ou presque, plutôt relâchée, plusieurs fleurs. Spathe ovoïde, avec un bec court, aux membranes blanches. Pédicelles presque égaux, bien souvent aussi longs que le périanthe, vert pâle, lisses, bractéolés à leur base ; bractées nombreuses, linéaires ou filiformes, à membrane blanche. Périanthe ovoïde, s'ouvrant indistinctement en fleurs. Tépale obtus, arrondi vers l'apex ou légèrement entaillé au sommet, blanc-vert avec une très fine veine médiante verte ; tépales extérieurs largement lancéolés, longueur 4,0-4,5 mn, largeur 1,5 mn ; tépales intérieurs linéaires et lancéolés, longueur 4 mn, largeur 1 mn. Filaments blancs, 1,5 fois plus longs que le périanthe, avec des bords finement ciliés ; filaments extérieurs à base triangulaire, filiforme, arrondis vers l'apex ; filaments intérieurs ovoïde à la base, plus large que les filaments extérieurs, s'amincissant étroitement dans le dernier tiers pour former une pointe portant l'anthère, avec une dent de chaque côté à peu près aussi longue que l'anthère centrale. Anthères roses. Pistil linéaire, jusqu'à 2 mn de longeur, entouré du périanthe. Capsule de 4 mn de long, 5 mn de diamètre ; valves rugueuses lorsque sèches, lisses. Graines comprimées, noires.

Phénologie : floraison en juin et juillet, fruits en juillet.

Écologie : L'espèce se retrouve sur des dépôts de grès de diverses sortes, entre 900 et 1500 mètres d'altitude au dessus de la mer.

Répartition : Figure 3. Kirghizstan, Ouzbékistan (contreforts et montagnes basses entourant la vallée de Ferghana. (Vvedensky 1971, Khassanov 2017)

TABLE DES ILLUSTRATIONS

Index des illustrations

TABLE DES MATIÈRES