REVUE
DE LA SOCIÉTÉ
DE PHILOSOPHIE
DES SCIENCES

Vol 8 N°2 2021

http://dx.doi.org/10.20416/LSRSPS.V8I2.4

Anouk Barberousse

BIODIVERSITY DATABANKS AND SCIENTIFIC EXPLORATION

SOCIÉTÉ DE PHILOSOPHIE DES SCIENCES (SPS)
École normale supérieure
45, rue d'Ulm
75005 Paris
www.sps-philoscience.org

latosensu
REVUE
DE LA SOCIÉTÉ
DE PHILOSOPHIE
DES SCIENCES

Anouk Barberousse

Biodiversity databanks and scientific exploration

Sommaire

For several decades now, biologists have been developing digital databanks, which are remarkable scientific instruments allowing scientists to accelerate the development of biological knowledge. From the beginnings of the Human Genome Project (HGP) onwards, genetic databanks have been a major component of current biological knowledge, and biodiversity databanks have also been developed in the wake of the HGP. The purpose of this paper is to identify the specific features of biodiversity data and databanks, and to point out their contribution to biodiversity knowledge.

Keywords: biodiversity, data, databanks, knowledge of biodiversity

1. Introduction

As soon as biologists became aware that climatic change and human action threaten not only some species but biodiversity as a whole, they began searching for new ways of developing biodiversity knowledge. Among these, building up biodiversity databanks has proven successful, as indicated by the growing number of such databanks. The words "biodiversity databank" are mainly used by professionals, namely by those who develop these databanks,[1] although a few historians, philosophers of science, and science studies scholars[2] have already begun to address this topic. As most developers of biodiversity databanks view biodiversity through taxonomic lenses, biodiversity databanks are usually filled with taxonomic data, progressively integrated with other types of data. They allow for the development of global knowledge of biodiversity in ways that call for epistemological analysis.

On the one hand, biodiversity databanks share some common purposes with other databanks in the life sciences, like genetic databanks; on the other hand, they take part in the development of new fields of research devoted to biodiversity. These two features will shape my inquiry in this paper, whose goal is to argue that biodiversity databanks allow for the exploration of biodiversity, a scientific endeavor whose features will be presented below. The first biodiversity databanks were developed at the same time as the first genetic databanks, and the success of the Human Genome Project (HGP) was responsible of transforming the meaning and potential of biological databanks in general, so that biodiversity databanks are now growing in the wake of the HGP, in the post-genomic era. The HGP has indeed provided all biological communities with standards of data sharing (Maxson Jones et al. 2018). The development of biodiversity databanks has also benefited from model organism databanks (Ankeny and Leonelli 2015). As emphasized by Maxson Jones and co-authors, "One crucial impetus for databases in the life sciences, alongside the persistence of natural historical practices and the FOSS [free and open-source software] movement, was the rise of model organism 'communities', such as those centered on fruit fly, yeast, mouse, and the nematode worm." (2018, p. 702). The HGP (and its -omic follow-up) and model organism databanks both provide a rich context (as analyzed in Leonelli 2016) through which to compare biodiversity databanks.

[1] *See, for instance, Beckett et al. (2020); Peterson et al. (2010); Hortal et al. (2015); Franz and Thau (2010).*
[2] *Biodiversity databanks have already been analyzed by a few scholars, from a historical (Strasser 2012, 2019) or science studies perspective (Bowker 2000; Devictor and Bensaude-Vincent 2016). The latter have mainly focused on the Foucauldian relation between classifications and political power on the one hand, and on the interplay between science and technology on the other, relying on the concept of technoscience. I will refrain from taking a stand on these themes.*

The notion of biodiversity may, in itself, be an object of philosophical inquiry, but instead of undertaking such an inquiry, I focus on the role of biodiversity databanks in this new scientific field. Coined in the 1980s as an abbreviated form for “biological diversity”, the term “biodiversity” has become so successful that it has acquired new dimensions of meaning. Besides referring to the diverse manifestations of life everywhere on the planet (whatever the temporal and spatial scale), it includes pragmatic concerns for conservation. This is why many discussions concern themselves with the best way to measure biodiversity. The topic of this paper is slightly different, however, and is oriented towards the means by which our *knowledge* of biodiversity is increased. Databanks are among these; they foster constitution of standards for biodiversity data, which, in turn, have a great unifying power over biological knowledge hitherto patchy and incomplete.

In order to identify the specific features of biodiversity databanks, I firstly investigate the transition from natural history to contemporary biodiversity knowledge and argue that biodiversity knowledge, as implemented in databanks, combines theoretical knowledge with scientific exploration in a new way. I then proceed by putting forward a (hopefully) ordered analysis of the constitution and organization of some biodiversity databanks. Lastly, I argue that the challenge of scientific integration, namely the main challenge for development of biodiversity knowledge via databanks, can be overcome when sufficient attention is paid to the methodology of scientific exploration.

2. From natural history to biodiversity knowledge

In the context of the current biodiversity crisis – whose symptoms are extinction on a massive scale and the degradation of many ecosystems – a major feature of biodiversity conservation endeavors is that they rely on poor knowledge of biodiversity and ecosystems. Biodiversity knowledge is thus caught up in a rather dramatic epistemic race against species extinction. The working hypothesis of the present paper is that to engage in this race, researchers and conservationists will, in the future, rely on the continuous development of databanks. These will provide invaluable information which will enable them to develop biodiversity knowledge. This working hypothesis will firstly be illustrated by emphasizing the link between current biodiversity databanks and natural history as it was practiced before their development at the end of the twentieth century.

Today, biodiversity databanks belong to different categories, defined by the nature of the data they include: taxonomic, genetic, geographical or various combinations thereof. Most taxonomic databanks are closely connected with collections of natural history in national or local museums. This is not the case for genetic databanks, although interconnections are not absent, as illustrated below. In this section, I focus on the material and scientific links between taxonomic databanks and collections of natural history because these links shed light on the historical continuity between current databanks and previous forms of biodiversity knowledge, as emphasized by Bruno Strasser. This will allow us to address the question of how to locate biodiversity databanks on the methodological map that features genetic and -omic data on the experimental side and natural history on the exploratory side. Scientific exploration has not been discussed very much within the philosophy of science, except for the role of some experiments in high-energy physics (Franklin and Perovic 2019; Steinle 1997). It is, however, the main method for learning anything new about biodiversity, along with evolutionary hypotheses, precisely because of the heterogeneity of the components of biodiversity that will be focused on in Section 2.3.

2.1 Natural history as big science

I firstly emphasize that despite biodiversity knowledge being obviously rooted in taxonomy, it does not solely consist of taxonomic knowledge. Whereas taxonomic identification remains the main channel through which biodiversity knowledge is produced, most biodiversity databanks also contain genetic, ecological, physiological and geographical information.[3] Thus, the links that I am going to assess between biodiversity databanks and collections of natural history do not only pertain to taxonomy but also to the whole of biodiversity knowledge.

Natural history, systematic exploration and description of the living world on a scientific basis, was a major field within early modern science, and big data played a part in this (Strasser 2012; Müller-Wille and Charmantier 2012). Scholars gathered specimens and descriptions thereof (including descriptions of the specimens’ environments) and stored these in large collections and scientific publications. The scientific (and political) importance of such collections was soon rec-

[3] It is important to distinguish between two features of items in biodiversity databanks: firstly, the nature of the involved information (genetic, taxonomic and ecological etc.); and secondly, the way this information has been established as reliable information. For instance, taxonomic information may well derive from genetic analysis; however, it will be classified as taxonomic if it pertains to species identification.

ognized by national governments, so that national collections were organized in such a way that all this knowledge was preserved and made accessible to scholars from all over the world. Although we currently associate big data with digital data, it is worth reiterating that collections of natural history are an early manifestation of the concept, accompanied by well-known practical questions of management and organization. Indeed, natural history has had to face the problem of organizing large amounts of data, establishing international standards and rendering specimens in collections accessible to scholars.

Today, we tend to see specimens in natural history collections as sources of data rather than as data themselves because they provide researchers with taxonomic, genetic, geographical and all sorts of -omic data.[4] As such, specimens form a link between biodiversity knowledge as it was exemplified before the realm of digital data and current digital databanks. The link between natural history and biodiversity databanks is not only material, though, since the contributions to knowledge provided by collections and databanks, respectively, share a common structure. Far from being purely descriptive, they are shaped by theoretical hypotheses. This point is worth expanding upon in some detail.

2.2 The role of hypotheses in natural history

I will firstly focus on taxonomy, the science of classification of living beings. Its relationship with underlying theoretical hypotheses is both ancient and complex. Taxonomy did indeed undergo a revolutionary theoretical change following the Darwinian revolution. In the eighteenth century, the principles of classification (stating the reasons why some specimens are considered as belonging to the same category) were determined by a fixist view of the living world. In contrast, today, they are grounded in evolutionary theory. Thus, taxonomic knowledge, which is implemented in the material organization of collections of natural history, decisively depends on underlying theories. In the same way, digital databanks implement current biological knowledge, be it taxonomic or genetic. This knowledge is mostly implemented in the design and adoption of standards that allow for comparisons: “An overriding concern among data-driven sciences, past and present, has been the production and enforcement of standards. Because comparative approaches are so crucial to data-driven sciences, the uniformity of the data has been essential.” (Strasser 2012, p. 86). The standards can only be designed by relying on generally agreed-upon hypotheses.

The above emphasis on the theoretical background of biodiversity databanks might seem surprising compared with the common view that databanks are a paradigmatic outgrowth of data-driven science, which is traditionally opposed to theory-based science. However, this usual association of natural history and biological databanks on the one hand, and data-driven science on the other hand, is not as justified as it initially seems. Bruno Strasser rightly emphasizes the following points:

“Natural history had been ‘data-driven’ for many centuries before the proponents of post-genomics approaches and systems biology began to claim the radical novelty of their methods. […] [M]any of what are claimed as novel features of contemporary data-driven science have parallels among earlier natural history practices. However[…], there are nonetheless important differences between past and present data-driven sciences. Most significantly, much of contemporary biomedical research represents a new hybrid of naturalist and experimentalist approaches. Today’s databases are as important to the experimentalists as museums were (and are) to the naturalists. Combining the data-driven and the hypothesis-driven, the comparative and the exemplary, the experimental and natural historical, current life sciences seem indeed headed in a new direction.” (Strasser 2012, p. 87)

Before further examining the “new direction” of current life sciences, as suggested by Strasser, I will introduce a brief comment about the role of theoretical hypotheses in biodiversity databanks. Stevens (2013), Bowker (2000), and Devictor and Bensaude-Vincent (2016) all tend to contrast datafication with reliance on hypotheses and theories. For instance, Stevens (2013) claims that the use of databanks is tantamount to giving up hypothesis and theory testing via observations and experiment, and Bowker (2000) insists that there is a “disarticulation” between data accumulation and knowledge production. On the other hand, Callebaut (2012) points to the theory-ladenness of data in databanks, as data are the result of modeling and intersubjective work. I emphasize that whereas the word “datafication” describes an important feature of the recent scientific endeavor with respect to biodiversity, it should not obscure the fact that biodiversity data are mostly pieces of patiently elaborated, qualitative knowledge rather than approximate, quantitative indicators.

[4] This setting is similar to that analyzed by C. Wylie (2018) in the case of fossils. Researchers face the same type of theory-ladenness and the weight of their theoretical decisions in both cases.

2.3 The "new direction" of scientific methodology: combining theoretical knowledge and scientific exploration

In light of the above emphasis on the continuity between natural history and current biodiversity databanks, I will now try to identify the main features of the "new direction" of scientific methodology that Strasser identifies, which combines theory-based and exploratory aspects. Exploration, in this domain of inquiry, is both data-driven and conditioned by a rich background of already established knowledge and agreed-upon hypotheses. The point here is that in this case, "data-driven" is not opposed to "theory-based", as the very establishment of secured data relies upon agreement on the hypotheses that are instantiated in the standards of data validation. A comparison with HGP-associated projects may be useful. As pointed out by Waters (2004), on the one hand, they rely on the already established knowledge that allows for gene sequencing and genetic data processing, but on the other hand, they are characterized by the will to practice sequencing without respect to particular hypotheses pertaining to the relations between genotype and phenotype which made it possible to, e.g., determine when and where miRNAs are produced. The latter result could not have been established without strict standards, themselves relying on theoretical knowledge.

In view of the above, it appears that the link between collections of natural history and biodiversity databanks sheds light on the epistemic aims of the latter. Natural history, as recalled above, is often conceived of as scientific exploration of the living world and as mostly relying on observation vs. experimentation, thus inducing a commonly adopted association between exploration and observation. In contrast, genetic data, as provided by strongly standardized bench techniques, are associated with experimentation vs. observation.

It might be thought, at first sight, that the exploratory capacity of biodiversity databanks is an effect of the way the quest for knowledge is handled in this domain. However, this exploratory capacity is better viewed as being dictated by the main features of biodiversity itself (vs. our means to know about it). Biodiversity has indeed been shaped by billions of years of evolutionary contingency, as much as by natural selection and genetic drift. This feature determines the specific features of scientific knowledge in this domain. The development of scientific knowledge is governed by both the search for regularities (as in physics, physiology and some parts of ecology) and by the identification of singular facts that have important evolutionary effects. Without taking contingent events into account, biodiversity knowledge cannot be considered complete. Thus, the structure of biodiversity knowledge is different from, e.g., the structure of physical knowledge, as the search for quantitative regularities is not central. Now, quantitative regularities are the most efficient way to organize a field of knowledge and provide it with architectural features. In contrast, with biodiversity being mainly shaped by evolutionary contingency, exploration (as opposed to the search for quantitative regularities) is the best method to acquire knowledge.[5]

Focusing on other life science disciplines, namely those identifying the functions of key biological molecules, Richard Burian also emphasizes the specific features of the domains that have been shaped by evolutionary contingency:

"[C]urrently, no systems for generating general hypotheses and no bodies of fundamental biological or chemical theory, supplemented by appropriate boundary conditions plus general background knowledge, are able to predict both in general and in detail genotype-phenotype relations, or structure-function relations for wide ranges of important biological molecules. A mixture of empirical, specialized theoretical, computational, and 'discovery methods' are required for these major tasks." (2007)

Burian completes his analysis of these essentially historical domains (in the sense that historicity cannot be set aside when scientific understanding is at stake) by recalling that "discovery methods require major instrumental and computational resources and yield very large quantities of data", to which we may add that in order to be useful, these data have to be collected, standardized and organized. Databanks are precisely the best-known way to preserve, organize and share the results of exploration. As they are influenced by evolutionary contingency, these results have to be processed in a way that makes them intelligible and useful for researchers and conservationists. The principles governing the organization of data in databanks are meant to answer these requirements, as we shall see in the next section.

[5] An anonymous reviewer, whom I warmly thank for this remark, suggested that biodiversity knowledge might be similar to historical knowledge (meaning knowledge of human history) because of the strong influence of contingency. However, historians usually do not use databanks in the same way as biologists do. The comparison is less straightforward than it seems, since economists, particularly historians of economics, use databanks and look for combined effects of regularities and contingent events in the same way as biologists do. Historians may generalize the use of databanks in the close future.

3. What are biodiversity data and how are they organized within databanks?

As mentioned above, biodiversity databanks are filled with a great variety of data: genetic sequences, complete genomes, taxonomic descriptions, taxonomic revisions, species occurrence data, physico-chemical measurement results, and so on. These data are obtained via a large variety of scientific procedures as well: genetic sequencing and -omic studies; taxonomic inventories; assessments of the state of biodiversity in a given area; curation and development of collections of natural history; and transformation of data stemming from other databanks. In this section, I firstly argue that besides the way they are produced, scientific data should also be analyzed according to their epistemic function – namely, becoming elements to be relied upon in further inquiry. In order to fulfill this function, data have to be validated. In Section 3.2, I examine how the various types of biodiversity data are validated. Lastly, in Section 3.3, I give examples of the different ways in which biodiversity data are organized in databanks and argue that although there are no centralized, standardized principles of organization, there is, nevertheless, an ongoing effort to guarantee inter-accessibility and interoperability, despite the variety of organizational principles.

3.1 The two sides of data

In order to assess how biodiversity databanks take part in the scientific exploration of biodiversity, it is important to recall that the concept of data may be analyzed from different points of view, each of which determines some specific aspect of the contribution of databanks to scientific knowledge. Firstly, the category of data itself is strongly historical and malleable, as emphasized by Maxson Jones, Ankeny and Cook-Deegan (2018, note 6, p. 698). Secondly, types of data might be distinguished either based on their production mode or on their epistemic functions. Let me now develop these two options and their respective implications. When types of data are distinguished based on their production mode, then observation data are distinguished from data obtained by experimental procedures. In this case, some taxonomic data (those coming, e.g., from morphological observation and comparison) are to be contrasted with genetic data, as they result from genetic sequencing procedures. On the other hand, when types of data are distinguished relative to their epistemic functions, other features of data are taken into account. Thus, some data are used as a means to obtain other data. For instance, in climatology, measurement results are often aggregated in order to obtain averaged data. In contrast, other data are used within a process of hypothesis testing; some are used as the basis for a search for patterns or regularities; lastly, we must also contend with the fact that the usefulness of yet more data is dependent on the reliability and security of previously obtained data. For instance, genetic data may be used to both test a selectionist hypothesis and complete our exploration of the diversity of genetic determinants for some phenotypic trait. In this section, I will follow functional data analysis in order to assess the organizational principles of biodiversity databanks and their potential for the scientific exploration of biodiversity.

When the epistemic functions (vs. origin) of data are pointed out, the importance of validation procedures that allow for the fulfillment of these functions is clear enough. These validation procedures, attesting the reliability of data included in the databanks, are the means for epistemic control, which is itself a condition of the development of sound knowledge. “Control” here means the capacity to assess (i) that data have been validated, (ii) how, and (iii) by whom. It goes beyond mere validation and is linked with traceability. Let me emphasize that epistemic control comes with data openness, however counterintuitive this claim might seem at first sight (a point also emphasized by Leonelli 2016 but with a lesser emphasis on the epistemic component of the control). On the one hand, data openness might appear to be a threat to reliability, as in the case of re-use of data by researchers who did not carry out the validation procedures themselves. But on the other hand, data openness is a precondition for epistemic control, since it is not only necessary that data be validated, but also that researchers be able to access the validation procedures (included in the metadata) *and* be able to correct the results thereof. as the case may be.

3.2 Validation of biodiversity data

As for biodiversity data, validation can be sought through different processes. For instance, validation of genetic data depends on the quality of each step of the sequencing process, but also on the care with which the resulting gel is analyzed and the hypothesized sequence transferred on a digital medium. In contrast, validation of taxonomic data depends on iteration of different steps composing integrative taxonomy, from morphological and genetic comparisons to inclusion of more specimens in the original dataset (see, for instance, Pante et al. 2014). The variety of processes through which validation of biodiversity data is performed constitutes an obstacle to scientific integration of data, as discussed in Section 4.

Let me now illustrate the variety of processes through which biodiversity data are validated, giving an example and comparing more precisely the validation procedures of genetic sequences on the one hand, and taxonomic descriptions on the other hand. They share the following features: Firstly, neither kind of data is easy to get and both require delicate control procedures; secondly, both rely on important amounts of already established biological knowledge. Thus, a genetic sequence is obtained at the end of a biochemical experiment (briefly presented above). The result of this experiment becomes the object of an interpretive judgment, delivered by a competent scientist, about the nature of revealed nucleotides and their order. I now turn to the users of the sequences and consider researchers who look for a genetic sequence in a databank. If they want to assess the reliability of a sequence they find, they have to rely on metadata, namely pieces of information that allow the validity of the sequence to be checked, for instance, the scientific paper in which it was first published. In the same way, a taxonomic description is the result of a complex inferential process, including (as in the genetic case) assessment of various hypotheses. The current rate of taxonomic revisions reveals how difficult this process is.[6] Thus, although their acquisition and validation processes are utterly different, genetic sequences and taxonomic descriptions share important features when considered as items within databanks. Both are attested by the series of robust collective processes resulting in their validation. These processes allow researchers and conservationists to achieve two important goals: (i) to use them as reliable data in their quest for further knowledge or to guide their conservation actions; and (ii) *when doing so*, and to the extent possible, to bracket all the previous elements that have contributed to their validation.[7] In other words, acceptance of genetic sequences and taxonomic descriptions as rightly belonging to biodiversity data means that, once validated, they can be used without further caution.[8]

3.3 Organizing biodiversity data

At the beginning of Section 3, I have presented the common features of biodiversity data via the double example of genetic and taxonomic data. These common features include the fact that most biodiversity data are by no means brute data but rather pieces of knowledge that have been validated by processes requiring reliance on theories and hypotheses. I now turn to the organization of biodiversity databanks. As with all human artifacts, biodiversity databanks are subject to tension between the desire to build them up according to rational and scientifically well-grounded principles, and the reality of existing and evolving technology, sometimes burdened by the necessity to transform artifacts from the past. In the case of digital databanks, transformations are often necessitated when new information systems are implemented in the institution that maintains the databanks. Biodiversity databanks are subject to another constraint as well, due to their object, namely biodiversity. As biodiversity extends throughout the entire planet, biodiversity databanks are indeed subject to what may be called a globality constraint. Unless specifically devoted to one geographical area, they should include data from everywhere on the planet.

Most biodiversity databanks are organized by the following principle: One class of data is selected as the principal one, and the others are linked with it, as secondary classes. For instance, the principal class of data within a databank may be genetic, taxonomic or geographical, or else it may be constituted of journal articles, pictures, entire genomes or lists of ecological traits. These classes of data are not exclusively identified by the type of scientific information they contain (like geographical, genetic or taxonomic information), but rather by a combination of the type of information *and* the format of the vehicle by which it is communicated (short text, scientific paper, image, genetic sequence, etc.). The principal class of data in a given databank allows its designers to define the center of the databank and to distinguish it from its periphery. (This metaphorical use of "center" and "periphery" is meant to bring some descriptive order to the analysis of biodiversity databanks in order to allow for comparisons, in spite of the differences prevailing in their origin and institutional aspects.) The diversity of the classes of data chosen as defining the centers of databanks is an obstacle to establishing information flows in biodiversity databanks. It also hinders interoperability of data, namely their capacity to be used efficiently in different databanks.

Let me now turn to some examples. I use them to illustrate my notion of "classes of data" and the distinction between the center and the periphery of a databank. They illustrate the unifying power of the standards upon which their international usability is based.

[6] *The centrality of taxonomic revision in biological classification is often overlooked by science studies approaches, which often view classifications as disconnected from the production of biological knowledge.*

[7] *Leonelli (2016) argues that tacit knowledge and familiarity with the objects described in the databases are necessary to make good use of these databases. What I want to emphasize here is that when proceeding to undertake further scientific inquiry, data should be relied upon without too much concern; otherwise, their users are simply blocked in their enterprise. Familiarity does not necessarily involve detailed knowledge or control of the validation process. I thank an anonymous reviewer who helped me to make this point more precise.*

[8] *This does not mean that databanks become hypothesis-free by the same token.*

I begin with the *Global Biodiversity Information Facility* (GBIF, https://www.gbif.org/), "an international network and research infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth" (https://www.gbif.org/what-is-gbif). Because it results from international working groups aiming to define international standards for the study of biodiversity, the GBIF "provides data-holding institutions around the world with common standards and open-source tools that enable them to share information about where and when species have been recorded" (ibid.) – namely, to deposit and manage occurrence data (as named by *the Darwin Core Standard*; see below).

The *Ocean Biogeographic Information System* (OBIS, http://iobis.org/) belongs to the same type of biodiversity databanks as the GBIF. It is centered on species occurrence data, combined with geographical information. This is a good example of a global, open-access databank that is easily connectable with marine databanks centered on taxonomic data.

The *World Register of Marine Species* (WoRMS, http://www.marinespecies.org) is such a databank, which establishes links between original species names (names that were given when the species was first identified) and current names. This allows taxonomists to keep track of taxonomic revisions, which, as we have seen, are an important element of the way in which biodiversity knowledge develops. Thus, WoRMS contains a list of names of marine organisms, including information on synonymy: "The system not only allows the storage of accepted and unaccepted names, but it also documents the relationship between names. This makes it a very powerful tool for taxonomic quality control, and also allows the linking of different pieces of information through scientific names." (http://www.marinespecies.org/about.php#what_is_worms).

Another element of taxonomy-centered databanks is worth mentioning, namely the existence of the *Darwin Core Standard* (DwC), which "offers a stable, straightforward and flexible framework for compiling biodiversity data from varied and variable sources" (https://www.gbif.org/darwin-core). More concretely, "Depending on how much information the source data contains—and how much they wish to share—publishers can create a Darwin Core Archive with one of three cores:

- a Taxon core, which lists a set of species, typically coming from the same region or sharing common characteristics

- an Occurrence core, which lists a set of times and locations at which particular species have been recorded

- an Event core, which lists field studies (including the protocols used, the sample size, and the location for each)." *(ibid.)*

Among other taxonomy-centered databanks, the *Barcoding of Life Data System* (BoLD, http://www.boldsystems.org/) combines genetic and taxonomic information and links genetic sequences with specimens gathered in collections of natural history. Its functioning is thus utterly different from the functioning of the most famous genetic databanks, namely GenBank (https://www.ncbi.nlm.nih.gov/genbank/), which is "the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences". It associates a genetic sequence with a published scientific paper, but not a specimen.

The history of GenBank has already given rise to historical papers shedding much light on the way that standardization, a huge challenge in the late 1990s and early 2000s, has been achieved (see, for instance, Maxson Jones et al. 2018). The goal of this paper is "to document the reciprocal constructions of data accumulation, data sharing, and science policy as molecular biology transitioned from its classical era to genomics" (note 6, p. 698). Whereas science policy is absent from the present paper, it is worth emphasizing that it has played a role in development of the BoLD. The kind of historical work that had shed light on GenBank is still to be done for biodiversity databanks (although an initial step has been taken in this direction by Devictor and Bensaude-Vincent 2016). This should reveal much about the actual way that standardization, scientific agreement and cooperation obtain[9] in the tense context of the biodiversity crisis.

My last example will be *Marine Species Traits* (http://www.marinespecies.org/traits/), whose aim is to connect ecological, geographical and taxonomic data in order to describe species patterns and the underlying processes explaining these patterns, as this "is essential to assess the status and future evolution of marine ecosystems" (http://www.marinespecies.org/traits/). Marine Species Traits is an interesting example of a current attempt to define new standards for biodiversity databanks, focusing on ecology, as ecological descriptions are poorly standardized (Costello et al. 2015). It is not even clear which features of the environment are to be selected in the ecological descriptions; the organisms are the only agreed-upon entities whose description has already been standardized, but even so their behavior is difficult to express by means of a common language. The relevant environment is also difficult to analyze into commonly agreed-upon traits, not to mention quantitative properties (a point also made by Devictor and Bensaude-Vincent 2016 and Bowker 2000).

By considering the above examples of biodiversity databanks, we can realize how difficult it is to establish reliable informa-

[9] For instance, it would be useful to study the "politics" of BoL data in the same way as Stevens has studied the "politics of sequence" (Stevens 2015).

tion flows throughout them and ensure interoperability. For sure, the establishment of international standards is *the* road towards such an achievement, but standards can only be devised when background knowledge is sufficiently stabilized. In the context of the biodiversity crisis and accompanying urgency, reliance on scientific knowledge has to be combined with other constraints. The connection between biodiversity databanks is thus a huge scientific challenge, necessitating identification of the conditions in which information flow is fruitful. I comment on this challenge in the next section.

4. The challenge of scientific integration

As pointed out in Section 3, biodiversity databanks can only contribute to biodiversity knowledge if they overcome at least some of the differences that stem from their having different principles of organization. The most serious challenge that researchers have to face when using and developing biodiversity databanks is indeed the integration of data which vary in origin and nature. In this section, I analyze the obstacles that these differences create for the organization of databanks. Before doing so, it is worth pointing out in what sense “scientific integration” is used in the context of the study of biodiversity databanks. Relative to the case studies assembled in the special issue on this topic edited by Ingo Brigandt (2013), scientific integration of biodiversity data is more on the side of data integration, as studied by Leonelli for the biology of plants (Leonelli 2013), than on the side of explanatory integration. As previously emphasized, biodiversity knowledge is only partly structured by the search for regularities and has to take an exploratory direction because of the importance of contingency in evolution. Therefore, there is no question of finding unified explanatory principles besides natural selection and genetic drift. However, data integration involves more than data accessibility and interoperability, as it is both guided by available theoretical knowledge (about evolution and genes) and fosters such knowledge; data integration is not just about data but also about knowledge development. Accordingly, in Section 4.1, I will examine the relationships between integration of biodiversity data and unification of biodiversity knowledge, and I will assess the prospects of the latter in spite of several obstacles. In Section 4.2, I will argue that connecting biodiversity databanks and collections of natural history is an efficient way to progress along the path to unification of biodiversity knowledge.

4.1 Integration of data; unification of knowledge

The first element to be pointed out in relation to the challenge of data integration is the scientific context within which it emerges, namely the current, incomplete state of biodiversity knowledge. On the one hand, biodiversity knowledge is in great need of unification because its very object – biodiversity – has, since the term was first coined, been conceived of as a global object whose unity, however tenuous, depends on there being only one known Earth on which there is a history of life. But on the other hand, biodiversity knowledge is utterly incomplete and patchy. Moreover, it suffers from a lack of theoretical integration because the various disciplines that, together, contribute to biodiversity knowledge are not themselves sufficiently unified, as I will now argue.

Ecology, taxonomy, phylogeny, population genetics, paleontology, biogeography and macro-ecology, which are the main disciplines contributing to biodiversity knowledge, all deal with the effects of evolutionary history. However, their characteristic scales (both spatial and temporal) are so heterogeneous that integration at the level of each discipline is already difficult. This is an obstacle to standardization of cross-disciplinary data, as we have seen with the example of *Marine Species Traits*. Standardization is facilitated when a common theoretical background is available, but it is hindered when it is not. Now, the above-mentioned disciplines are at least linked by evolutionary theory, but the link is rather weak and does not allow researchers to overcome the discrepancy of spatio-temporal scales.

Another obstacle to scientific integration at the level of each involved discipline contributing to biodiversity knowledge is that there is currently no consensus on the right unit of biodiversity. Genes, populations, species, communities, ecosystems and landscapes are all plausible candidates, implying very different conceptions of the dynamics of biodiversity transformation. Beyond disagreement about this question, there even seems to be a rather strong theoretical dissensus underlying it. Indeed, the various answers to the question of the right unit for biodiversity echo deep divergences in ecology and evolution about units of evolution and units of selection, and about the respective weights of selection and genetic drift etc. These are the main theoretical questions shaping the hoped-for articulation between ecology and evolution. In view of the current absence of consensus on these theoretical questions, it seems that there is no hope of reaching any agreement on units of biodiversity that would come from the theoretical side. A response to this observation may be that one purpose and distinguishing value of databanks is precisely to overcome theoretical disagreement and guarantee scientific integration by proceeding bottom-up, from data to theory. However, as we have seen, the "data" in the case of biodiversity are already complex pieces of knowledge, relying on a dense theoretical background. The bottom-up move thus brings its own problems and cannot be seen as an all-purpose problem-solver in this context.

Lastly, another tremendous obstacle to scientific integration of biodiversity data is worth mentioning. This has to do with the ratio between the amount of available data and the amount of data that would allow for satisfactory knowledge. This ratio is very small as available data fail to cover large geographical areas. This failure cannot be overcome by resorting to generalizations, and this is due to the two reasons that were emphasized above: historical contingency and the fragmentation of disciplines.

4.2 Connection between databanks and collections of natural history

Against this background, a secure connection between digital databanks and collections of natural history appears as a possible pathway towards meeting the challenge of scientific integration of biodiversity data. The first point in favor of this connection is that it is the only way to avoid the burden of taxonomic uncertainty. This is why securing the link between taxonomic and genetic information and its material sources is so important. However, taxonomic uncertainty, which is certainly a major concern for taxonomists, is not usually seen by other biologists as a threat worth considering. These diverging assessments feature in the problem of integration of biodiversity data, which reveals itself at the level of scientific communities (in contrast with problems emerging from the nature of the data themselves). The point here is that biologists other than taxonomists consider species identification as given once and for all. However, species identification, as well as the constitution of taxonomic knowledge in general, is an ongoing process of hypothesis testing, rather than the final disclosure of what there is on Earth. The misinterpretation of taxonomy as a precondition of other disciplines rather than a partner in the process of knowledge building results in underestimating the importance of being able to check each step linking a genetic sequence with an organism. Ensuring the possibility of checking, however, is the only way to connect genetic sequences in their digital format with the ultimate, material targets of inquiry, namely organisms.

The second reason why the connection between digital databanks and natural history collections is an important condition of scientific integration of biodiversity data is that collections of natural history are epistemically powerful due to their age. Connection with collections of natural history thus anchors contemporary big data to the long-term history of scientific knowledge. Even though biological knowledge has been profoundly transformed by the successive revolutions of Darwinism, the discovery of DNA and the sequencing of complete genomes in the context of the HGP, one part of it has been surprisingly stable, namely operational taxonomy and nomenclature. Species are identified and given names according to methods that have not changed much since the eighteenth century, even though the set of characters used in species descriptions has been significantly enriched by the recent availability of genetic characters. The claim that taxonomy relies on stable practices might look controversial at first sight, mainly because the theoretical foundations of Linnaean taxonomy did not survive the Darwinian revolution and also because taxonomic revisions happen on a daily basis. However, despite theoretical change and taxonomic revisions, continuity is a remarkable feature of the practices of species identification. One reason for continuity is that the scientific community has been extremely careful in keeping track of taxonomic revisions: firstly, within scientific papers and monographs, and now, within databanks, as we have seen with the example of WoRMS. For non-taxonomists, information about the history of taxonomic revisions and the way species names transform (and the reasons for these transformations) look like dispensable technicalities; however, they should be viewed as a paradigmatic example of a rigorous scientific practice of knowledge management, conducted over the long term. Now, in spite of the recently coined term "biodiversity", the development of biodiversity knowledge is rather old; this is why older scientific practices should be taken into account when trying to develop it in the new, digital area. Digitalization is a facilitator of knowledge development, but it has to be combined with older scientific practices if the older epistemic efforts are not to be wasted.

I have just argued that taking into account the history of the development of biodiversity knowledge can contribute to the scientific integration of biological data. It implies being aware of the fine-grained methodological challenges that arise in the process. Building up biodiversity knowledge indeed consists of unifying heterogeneous pieces of knowledge whose degrees of reliability are variable. This not only involves managing large amounts of data but also establishing second-order links about standards of validation, as the data in biodiversity databanks are pieces of knowledge whose justifications are of various types. Thus, what is at stake is not only the *quantity* of data and their interoperability but also interoperability of standards of validation. The latter may be obtained by designing processes that will guarantee the same degree of reliability for all sets of data, to the extent that it is possible.

4.3 Scientific exploration

Let me now turn to the implications of the above discussion for the way biodiversity databanks contribute to the development of biodiversity knowledge, and come back to the theme of scientific exploration. I have argued that the transformation and dissemination of standards is a powerful way to develop biodiversity knowledge, although different from more classical ways like finding generalizations allowing for predictions. I now address the question of how these two ways of developing biodiversity knowledge compare and interact. Philosophers of science have mainly focused on knowledge generation based on the discovery of general truths about phenomena; however, producing true general statements is only one way to achieve scientific knowledge. Producing reliable data and organizing them within accessible and interconnected databanks is another. This can only occur if standards of reliability are openly discussed and established on an international basis, as emphasized above. Establishing standards is the first step towards achieving general knowledge. As such, it contributes to the methodology of scientific exploration. Within scientific methodology at large, exploration is often overlooked, however. But in order for the exploration to be scientifically fruitful, its outcomes have to be usable by all researchers taking part in the development of biodiversity knowledge. This, in turn, involves meeting standards of reliability that transform as knowledge expands and tends towards greater unification. “Datafication” is not something that is achieved once and for all; it develops (or should develop) at the same pace as scientific knowledge, generalizations, hypothesis testing and modeling etc.

I have to insist once more that all the above implies that integration of biodiversity data is not only a technical but also a scientific problem because it has to keep up with the continuous development of scientific knowledge. Transformation of scientific knowledge of biodiversity not only involves designing new hypotheses and a better connection with fundamental hypotheses about evolution and ecological processes, but also the transformation of validation criteria and the re-assessment of criteria defining what is a reliable piece of knowledge. As a result, scientific integration of biodiversity data is a never-ending dynamical process, whose end is, moreover, not fixed in advance.

A consequence of the dynamical character of scientific integration of biodiversity data is related to the above-mentioned question of how we should conceive of scientific exploration. Scientific exploration is not only a matter of going to some region and finding out what is there but, more importantly, of understanding how what one finds out relates to what is already known. This process is entirely different from the one in which things are discovered from scratch, but it is more about fitting what has been discovered into the web of existing knowledge. The metaphor of the web of existing knowledge is meant to refer to a set of already established facts and data but also hypotheses with attached degrees of plausibility. Some of these hypotheses will be given up and others will become knots, but the very nature of this “web of existing knowledge”, at any given time, is such that it will contain hypotheses whose prospects are uncertain.

The main upshot of this inquiry into the challenge of scientific integration of biodiversity data within interconnected databanks is that it allows us to identify a new domain within the study of scientific methodology. Scientific methodology should not be divided into hypothesis-driven and data-driven methods, but one should recognize instead the importance of the *process*, in the long run, of scientific exploration, which involves integration of heterogeneous data. The design and interconnection of databanks participate in this process.

5. Conclusion

I have firstly presented how knowledge of biodiversity has developed from old taxonomic practices associated with collections of natural history to current digitalized databanks, including genetic and other types of data. This has led me to emphasize that whereas the datafication of biodiversity has often been viewed as a process of cutting data off from theories and hypotheses, development of biodiversity databanks, on the contrary, participates in the dynamics of scientific, theory-based knowledge. This was true in the past, as taxonomy has been shaped by the process of taxonomic revision from its very beginning; and this is still true now that taxonomic knowledge is combined with other parts of biological knowledge. In Section 3, I have shown how the close connection between datafication of biodiversity and the necessary growth of biodiversity knowledge due to climatic urgency translates into the organization of databanks. I have distinguished between classes of data according to their relative positions within databanks (at the center vs. at the periphery) and have discussed how this organization, due to the variety of research domains that contribute to biodiversity knowledge, makes information flow and interoperability more complex. Lastly, I have addressed the main challenge facing biodiversity databanks, namely scientific integration. The fields contributing to biodiversity knowledge are so diverse and their theoretical foundations so dis-unified that operational integration, a precondition for interoperable datafication, is immensely difficult. Analyzing this challenge has led us to discuss how the new availability of biodiversity data transforms the interactions between hypothesis testing and exploratory research.

In sum, building up sound and reliable biodiversity databanks is a genuine scientific endeavor, as the future development of biodiversity knowledge is dependent upon these databanks. The scientific work that is devoted to their design and interoperability is a good example of a sound practice of scientific exploration, a type of scientific method that has immensely benefited from the digitalization of data management and processing.

ACKNOWLEDGMENTS

This paper has a history in a joint work with Sophie Bary and Sarah Samadi (Paris MNHN). I warmly thank the audience of SPS2018 in Nantes, as well as the anonymous referees, who have made many valuable suggestions.

REFERENCES

ANKENY, R.A. and LEONELLI, S. 2015. Valuing Data in Postgenomic Biology: How Data

Donation and Curation Practices Challenge the Scientific Publication System. In RICHARDSON, Sarah S. andSTEVENS, Hallam (eds.). Postgenomics: Perspectives on Biology After the Genome. Durham, NC and London, UK: Duke University Press. 126–149. Lien

Sterner, Beckett W., Gilbert, Edward .E., and Franz, Nico.M. Decentralized but globally coordinated biodiversity data. Frontiers in Big Data, Oct. 2020. Lien

Bowker, Geoffrey C. 2000. Biodiversity datadiversity. Social Studies of Science, 30: 643–683. Lien

Brigandt, Ingo., ed. 2013. Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity. Special section of Studies in History and Philosophy of Biological and Biomedical Sciences (Volume 44, Issue 4, Part A, pp. 461–571). Lien

Burian, Richard M. 1997. Exploratory experimentation and the role of histochemical techniques in the work of Jean Brachet, 1938–1952. History and Philosophy of the Life Sciences, 19: 27–45. Lien

Burian, Richard M. 2007. On microRNA and the Need for Exploratory Experimentation in Post-Genomic Molecular Biology. History and Philosophy of the Life Sciences, 29(3). Lien

Callebaut, Werner. 2012. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1): 69–80. Lien

Costello, Mark John, Claus, Simon, Dekeyzer, Stefanie, Vandepitte, Leen, Tuama, Éamonn Ó, Lear, Dan, Tyler-Walters, Harvey. 2015. Biological and ecological traits of marine species. Peer J., 3: e1201. Lien

Devictor, Vincent, Bensaude-Vincent, Bernadette. 2016. From ecological records to big data: the invention of global biodiversity. History and Philosophy of the Life Sciences, 38: 13. Lien

Franklin, Allan, Perovic, Slobodan. Experiment in Physics, The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), ed. Edward N. Zalta. Lien

FRANZ, Nico, THAU, David. 2011. Biological taxonomy and ontology development. Biodivers. Inform., 7: 45–66. Lien

Hortal, Joaquín , de Bello, Francesco, Diniz-Filho, José Alexandre F., Lewinsohn, Thomas M., Labo, Jorge M., Ladle,Richard J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst., 46: 523–549. Lien

Leonelli, Sabina. 2013. Integrating data to acquire new knowledge: Three modes of integration in plant science. Studies in History and Philosophy of Biological and Biomedical Sciences. Lien

Leonelli, Sabina. 2016. Data-Centric Biology: A Philosophical Study. Chicago, IL: University of Chicago Press.

Maxson Jones, Kathryn, Ankeny, Rachel A., Cook-Deegan, Robert. 2018. The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. Journal of the History of Biology, 51: 693–805. Lien

Müller-Wille, Staffan, Charmantier, Isabelle. 2012. Natural history and information overload: The case of Linnaeus. Studies in History and Philosophy of Biology and Biomedical Sciences, 43: 4–15. Lien

Pante, Eric, Schoelink, Charlotte, Puillandre, Nicolas. 2014. From Integrative Taxonomy to Species Description: One Step Beyond. Systematic Biology, 64(1): 152–160. Lien

Peterson, Andrew T., Knapp, Sandra, Guralnick, Robert, Soberón, Jorge, Holder, Mark T.. 2010. The big questions for biodiversity informatics. Syst. Biodivers., 8: 159–168. Lien

Steinle, Friedrich. 1997. Entering new fields: Exploratory uses of experimentation. Philosophy of Science, 4 Suppl.: S65–S74. Lien

Stevens, Hallam. 2013. Life out of sequence—A data-driven history of bioinformatics. Chicago: University of Chicago Press.

Stevens, Hallam. 2015. The Politics of Sequence: Data Sharing and the Open-Source Software Movement. Information & Culture, 50(4): 465–503. Lien

Strasser, Bruno J. 2011. The Experimenter's Museum. GenBank, Natural History, and the Moral

Economies of Biomedicine. Isis, 102: 60–96. Lien

Strasser, Bruno J. 2012. Data-driven sciences: From wonder cabinets to electronic databases. Studies in History and Philosophy of Biological and Biomedical Sciences, 43 (2012): 85–87. Lien

Strasser, Bruno J. 2019. Collecting Experiments: Making Big Data Biology. Chicago: University of Chicago.

Waters, C.K. 2004. What was classical genetics? Studies in History and Philosophy of Science, 35: 783–809. Lien

Wylie, Caitlin Donahue. 2019. Overcoming the underdetermination of specimens. Biology & Philosophy, 34: 24. Lien

HISTORIQUE
Article soumis le 20 juin 2019.
Version révisée soumise le 3 mars 2020.
Accepté le 15 février 2021.

SITE WEB DE LA REVUE
https://ojs.uclouvain.be/index.php/latosensu

ISSN 2295-8029
DOI http://dx.doi.org/10.20416/LSRSPS.V8I2.4

SOCIÉTÉ DE PHILOSOPHIE DES SCIENCES (SPS)
École normale supérieure
45, rue d'Ulm
75005 Paris
www.sps-philoscience.org

CONTACT ET COORDONNÉES :

Anouk Barberousse
UFR de Philosophie, Sorbonne Université, Faculté des Lettres, 1 rue Victor Cousin, 75005 Paris ;
anouk.barberousse@sorbonne-universite.fr