JRAI Journal of the Royal Anthropological Institute

3
The datafication of nature: data formations and new scales in natural history

Tahani Nadim *Museum für Naturkunde Berlin/Humboldt-Universität zu Berlin*

In this essay, I consider the scales and connections lost and gained as natural history adopts digital data infrastructures. On the basis of ongoing work in the Museum für Naturkunde Berlin, I track the relations between insect specimens and their material and digital informational ecologies. Using Latour's notion of the 'circulating reference', I follow the insect specimens as they make their way into taxonomies, databases, and digitization apparatuses. In focusing on human-data mediations in museum practices of ordering, describing, and distributing specimens, I show how the datafication of nature makes present conventionally dissociated contexts, including German colonialism. Proposing the concept of a data formation, I suggest that ethnographers have much to contribute in bringing forward the sociocultural and historical specificities and contingencies within data.

I'm sitting in the office of the head curator for the Lepidoptera (butterflies and moths) collection at the Museum für Naturkunde Berlin, one of the world's oldest and largest natural history museums, to talk about the relation between global data infrastructures and museum collections. Perched between us atop the table is an insect drawer, a wooden rectangular box with a glass lid that is filled with dozens of tiny moths neatly pinned in tight rows. An A4 sheet of paper lies on top of the drawer. It shows a phylogenetic tree diagram, consisting of branches, nodes, and leaves arranged in a circle that represent the evolutionary relationships between the moth species assembled in the drawer underneath. The moths' genetic data, derived from their legs, had just been received from a Dutch-based biotechnology company specializing in DNA sequencing services. Based on similarity analysis, the sequences were grouped and subsequently translated into the diagram, where the lengths of the tree's branches are proportional to the amount of character change, thus indicating the relationship level between the moth species.

We are talking about the speed of taxonomic research, which, together with the care and management of collections, constitutes the curator's main work. Taxonomic research is concerned with the description, identification, and naming of species, and

Journal of the Royal Anthropological Institute (N.S.) **27**, 62-75

natural history museums form a central site for this pursuit in providing a home for experts and materials, most importantly specimens. Nowadays also referred to as biodiversity discovery, taxonomic work has a long history, usually said to originate with Carl Linnaeus (1707-78), whose system for naming organisms remains in place to this day. The artefacts furnishing the curator's office bear witness to the ongoing traditions still directing much of this work: a microscope; system trays (small cardboard boxes containing specimens); pincers and pens; more wooden drawers filled with pinned butterflies; a desktop computer; boxes overflowing with spare pins and labels. Book shelves line two walls, crammed with aged leather-bound volumes, papers, and journals. At the same time, the piece of paper bearing the tree diagram is indicative of how novel data-based technologies, including DNA sequencing, are changing species discovery, the material practices of taxonomic work, and its infrastructures and institutions. The curator tells me that the real bottleneck in scaling up taxonomy and fully realizing a data-driven biodiversity discovery is the tenuous, often non-existent connection between molecular (genetic) data and the published species names. 'We have', he emphasizes, '250 year's worth of species names' derived from identifications based on morphological traits such as colour, size, and shape. Now these need to be matched to the unique genetic signatures for each species, so-called barcodes, that became attainable through genetic sequencing at the end of the twentieth century. And, indeed, the diagram bears no species for the moths, just so-called Barcode Index Numbers (BINs), denoting algorithmically derived clusters of barcode sequences assumed to represent distinct species.

What interests me in this essay are the scales and connections lost and gained as natural history becomes populated by BINs, barcodes, and digital data infrastructures. In the following, I explore the meeting of what the editors have called the 'data moment' and practices of natural history, including discovering, naming, and digitizing, in the Museum für Naturkunde Berlin. My aim is to track this conjuncture through the connections and disconnections between the specimen and its emergent informational ecology. In doing so, I advance the concept of 'data formation' to gather the heterogeneous, contingent, and not always explicit processes that make and remake data in the context of natural history collections as they seek to transform themselves into global data infrastructures. Guided by Stoler's use of the term 'imperial formations' (2008: 192), data formations refer to human-data mediations adjacent to and exceeding the formal processes of computerized programming and coding. I argue that thinking and handling data as data formation complicates expectations and assumptions associated with data (Kitchin 2014), including the abstraction and standardization of phenomena, while also prompting a rethink of the categorical differences between big and small, digital and physical, past and present. Specifically, I demonstrate how the decontextualization of museum specimens in the course of databasing, classifying, and digitizing always entails a recontextualization that does not so much dispel as distribute complexity. My concern here is also to make evident how data formations point to a datafication of nature that is characterized by novel scales for natural history.

With this text, I want to bring into conversation two sets of literature: the body of work generally concerned with what has become called datafication (boyd & Crawford 2012; Ruckenstein & Schüll 2017; van Dijck 2014) and its sociopolitical effects; and social and cultural studies of science, more specifically, studies examining the lives of the objects and materials that go into the making of science and its claims (Franklin 2003; Knorr Cetina 1999; Latour 1999; Latour & Woolgar 1986). A specific analytical category

to emerge from the latter is the 'inscription', which describes the transformation of entities and phenomena into mobile traces such as diagrams, documents, or signals (Latour & Woolgar 1986). It is in the form of inscriptions that claims about the world can move beyond the field and the laboratory, get tested against and combined with other inscriptions, thus allowing the construction and manipulation of entities, from planetary boundaries to genetic barcodes. Inscriptions are not representations of the world but specifically materialized representations in the world that possess obduracy and agency and thus shape the world. For ethnographies of laboratories and scientific knowledge production, inscriptions and 'circulating references', which describe the contiguous translations of entities into signs, have served as central objects to follow and observe in action (Latour 1999). With expanding datafication, not only do these objects multiply, but their trajectories are ever harder to trace across machines, databases, software, and formats. Studies examining the proliferating incursions of data-driven quantification and governance into most domains of life afford data similar capacities to intervene in the world (Turnhout, De Lijster & Neves 2014). Here, too, entities and phenomena such as pollution (see Blacker, this volume) are transformed into inscriptions that can circulate, be aggregated, and become the target of policy interventions. Both sets of literatures are thus concerned with the power of data (and of inscriptions as data) to materialize and scale the world, and, furthermore, with the methodological challenges posed by the scale of data.

In the context of the data moment, the problem of scale is conventionally imagined through the tensions and trade-offs between big data and small data (boyd & Crawford 2012). As the editors of this special issue point out, this is usually where anthropology comes in, since 'small' here means empirically constructed thick data derived from local and situated engagements. Small data is regarded as close to bodies and the ground and steeped in context that can be neither vanquished nor abstracted without considerable information loss. And so the tension between big and small data reproduces a scalar problem common to both anthropology and natural history, namely the incongruity between 'taxonomic' and 'relational' precision (Wagner quoted in Strathern 2004: 121 n. 3). One could either hone in on the individual butterfly and describe it in all its details (taxonomic precision), or take in the entire drawer, consider the entire collection and all of the species' ecological encounters to describe its evolutionary development, environmental function, or similar relational complexes (relational precision). By proposing the figure of data formation for imagining and tracing data-human mediations in natural history, I want to move away from this dichotomy and suggest instead that data formations gain detailed shape as they also gain relationality. This is an urgent question as natural history collections are moving to comprehensively digitize and make their specimens available in global data infrastructures. In the course of this, decisions need to be taken as to what information to retain and at which level of detail, thus structuring access and inquiry in particular ways.

Taking my cues from anthropological and, more recently, science and technology studies engagements with scale and scaling, I want to examine scale as an empirical question and by doing so move away from the convention that sees big and small, the general and the specific, as mutually exclusive domains (Asdal 2020; Strathern 1995). The essay specifically attends to how data is imagined in natural history, past and present, by tracing and describing different yet connected moments of data formations: a database record and its specimen; the naming of a wasp; and the mass digitization of insects. Guided by actor-network theory's methodological proposition to 'follow the

actors' (Latour 1987), both human and nonhuman, I track inscriptions, references, and specimens through practices in the Museum für Naturkunde Berlin, which has been my institutional home and fieldsite since 2013. This following is based on participant observation, ongoing dialogues with colleagues, and an engagement with data-related developments through literature and observation.

Retracing references, returning complexities

At one point in my conversation with the Lepidoptera curator, we move to his computer so that he can show me the barcode sequences which have been derived from the moths and form the basis for the phylogenetic tree diagram. He opens the barcodes as a text file, revealing a set of sequences of As, Cs, Gs, and Ts, copies one of the sequences, and pastes it into the search function of the Barcode of Life Data System (BOLD) which he has called up in his browser. BOLD provides, among other things, a public database of reference sequences from vouchered specimens: that is, animals or plants preserved in institutional collections that serve as a verifiable and permanent record for the species. At first, no match is found, but a second sequence matches an Australian moth in the Australian National Insect Collection by over 90 per cent, allowing the curator to establish an informed connection between the moths in his office, the barcode, the valid species name, and the voucher specimen. According to Latour (1999), this traceability is a key condition for the efficacy of references to serve as both representative (of a species, environments, collections) and guarantor for downstream inscriptions such as publications. Given the plethora of historical and ongoing natural history collections, the proliferation of data portals like BOLD and data-based biodiversity discoveries like barcoding make these 'chains of transformation' (Latour 1999: 70) – that is, retraceable connections between specimens and references – more imperative since they multiply the specimen and its data. Retracing, however, might also proliferate relations.

I was searching through the Global Biodiversity Information Facility (GBIF), the largest online public database of biodiversity information, when I arrived at a small set of butterflies from the Museum für Naturkunde collected in Cameroon.[1] The thirty-nine database records appeared on my screen listed alphabetically by species name: *Anapisa aurantiaca*, *Automolis invaria*, *Balacra daphaena*, and so on. The entries shared geographical co-ordinates (4.3N, 9.1E), a date (1991 December), and the 'basis for the record' (Preserved specimen). They also comprised full taxonomic lineages, from Kingdom (Animalia) all the way to the species level. Clicking on the record for *Graphium fulleri*, the full database entry is called up. It is titled 'Graphium fulleri (Grose-Smith 1883)', indicating that Henley Grose-Smith (1833-1911), a British entomologist, was the first to identify the species and publish the name in 1883. It further contains the butterfly's common name (Riley's graphium), a not very detailed map of present-day Cameroon, as well as three photographs. The first shows the specimen from the top, the second shows it from the bottom, and the third depicts all the specimen's labels neatly arranged side by side. Curiously, the labels state the species name as '*Papilio sanganoides*', but this escapes my initial perusal. The photographs are followed by a table holding data relating to the record, the occurrence, identification, taxon, and location and a set of rows providing 'Other' data, in this case information on the record licence (Creative Commons 4.0). The specimen is marked as a holotype, which describes the name-giving specimen for a species (thus counting among the most valuable holdings of the collection). Like most holotypes in the collection, this

specimen, too, is accompanied by a tiny red label, captured in the photograph, bearing the word 'Type'.

In order to trace the reference of the database record back to the butterfly specimen, I meet the curator for the Lepidoptera collection. We make our way to the office of the collection care technicians, who are responsible for maintaining and preserving collection specimens. The collection rooms are cavernous, expansive yet dim, as cabinets and cupboards tower all the way to the ceiling, forming narrow alleys dotted with ladders and old-fashioned trolleys. Were it not for the odd piece of modern equipment, such as a vacuum cleaner or a freezer, a cursory view of the space would cast the observer back into the time when the building first opened in 1889. In addition to the species name, I had gleaned from GBIF the specimen's 'catalogue number' in the hope that the collection's catalogue might contain some contextual information which had not been integrated into the database record and which could tell me more about the circulation of the reference. Stacked in the technician's office, the collection catalogues are oversized leather-bound volumes of mildly yellowed pages bearing a double-spread grid that contains, in extremely neat handwriting, catalogue number, species name, location, collector, and, in rare cases, additional information such as details concerning acquisition. Unfortunately, as the technician remarks, the catalogue entries do not include dates, making it difficult to cross-reference with specimen labels and other documentation (such as accession books, inventories, auction catalogues, or annual reports). The number I had noted down for the specimen did not correspond to the specimen listed under the same number in the collection catalogue, prompting the technician to surmise that it probably matches the 'main catalogue', which is kept in the historical archives of the museum. This mismatch points to the specific history of the collection, which had initially begun as part of the zoological collection before this was divided into sub-collections corresponding to taxonomic groups (Mammals, Birds, Reptiles and Amphibians, Fishes, etc.). The technician grabs a slim brochure from behind her desk, the *Mitteilungen aus dem Zoologischen Museum in Berlin* (Notes from the Zoological Museum in Berlin) from 1904, which documents collection development. For the Lepidoptera collection, the report notes that most work consisted of preparing butterflies from the African colonies and that work on a general systematic catalogue for butterflies had started.

The curator and I continue on our search for the GBIF specimen, which takes us to the second floor of the building, where another part of the Lepidoptera collection is stored. Once more, we find ourselves between towering cabinets, each of which holds hundreds of specimens carefully pinned in insect drawers. At the back of one of the alleyways, we meet a volunteer, an elderly woman, carefully moving butterflies from old drawers into new ones. Volunteers are a common sight in the museum, providing indispensable help in caring for collections. In addition, by transferring specimens into new drawers and updating labels, they assist in the transformation of collections into modern information infrastructures. We move further down into the collection and finally arrive at a cabinet containing specimens collected on the African continent meant to hold the taxon *Graphium fulleri*. The curator slides out drawers, searching for the species while I call up the GBIF record on my smartphone to provide him with a visual for the butterfly. He pulls out a drawer, places it on a nearby table, opens it, and delicately removes a promising-looking butterfly. In order to check its identity, he removes the stack of labels pinned underneath it piece by piece. Despite the similarity, the information on the labels reveals this to be the wrong specimen and we return to the

cabinet, where another butterfly catches the curator's attention. Once more, however, we encounter a mismatch. While this certainly is the specimen in the photograph on GBIF – the markings are identical, as are the tiny oddities (missing antenna, crooked leg) – the taxonomic name on the label appears as '*Papilio foersterius*'. The curator looks again at the photographs provided on GBIF and inspects the labels there more closely, seeing, for the first time, the name given as *Papilio sanganoides*. This prompts him to speculate that while the database record lists the current valid name for the species (*Graphium fulleri*), all other names constitute synonyms, a product of multiple descriptions of the same species or of a change in classification or nomenclatural code.

One of the labels accompanying the museum specimen and featured on the photograph in GBIF bears the location of where the butterfly had been collected: 'S. Kamerun' (South Cameroon), 'Ngoko Sanga'. It also contains what presumably is the name of the collector ('Foerster'), who appears to have been honoured by having the species named after him (*Papilio foersterius*). It is likely that this refers to Oskar Foerster (1871-1910), a German colonial officer who had served in various expeditions in southern Cameroon while it was a colony of the German Empire (1884-1916) (Schnee 1920: 651). I had been told that the specimens from Cameroon listed in GBIF had most likely come to the museum in the late nineteenth or early twentieth century. This date range is consistent not only with the above-mentioned report from 1904, but also with the historical circumstances which saw German colonial troops engage in expeditions in southern Cameroon (Nghonda & Zacharie 2007). At the time, the region was the site of prolonged border negotiations and conflicts between German and French colonial powers. Expeditions combining military and scientific expertise were used to map territory and people, drawing boundaries and collecting specimens of local fauna and flora, which also served to naturalize colonized territories.[2]

Odd connections and disconnections accompany the database record once it is accessed not as a scientific reference but as a data formation. It makes evident that digital transformations inherit, and at times heighten, the problems that continue to haunt natural history, including its participation in imperial formations (Subramaniam 2014).

Names of transmission

When I joined the Museum für Naturkunde Berlin in 2013 on a post-doctoral fellowship as the museum's first social scientist, it was set to transform itself into a 'biodiversity discovery factory', as one key colleague put it. Laconically described by him as 'animals in, papers out', specimens would be delivered to and processed in the museum and turned into digital data about morphology and lineage, species abundance, distribution and trends, data about the state of ecosystems and biodiversity loss. This vision signals key aspirations undergirding the data moment, including automation and the seamless integration of disparate data, sites, and times, from historical specimens to genetic sequences, from museum collections to global databases. In retracing the database record, it has become evident that current digital data infrastructures remain deeply entwined with traditional taxonomic work and institutional histories. The vision of the biodiversity discovery factory imagined the transformation of natural history into a data-driven science and natural history museums as the centre of calculation and datafication. In fact, as colonial collections demonstrate, natural history museums have long been key to rendering worlds and their inhabitants into (digital) data. Not only does natural history constitute a vast

archive of data, but also its technologies for sorting, storing, relating, and distributing data persist in current data efforts that go beyond the 'traditional' remit of natural history (although, of course, this remit has never been strictly domained, as feminist and postcolonial scholars have demonstrated repeatedly). Despite the data diversity and unstable taxonomies, it is important to note that these remain hegemonic descriptions of the world. The datafication of nature, and, by extension, the data moment with which the essays in this volume grapple, are outcomes and continuations of specific histories that endure, even when technologies, economies, and routines change. The datafication of nature thus compels a reckoning with the tenacity of natural history's modes of ordering, especially its naming practices. Names are scalar devices in that they allow a species to emerge from a specimen, thus affording a switch of perspective from the taxonomic to the relational. In the following, I retrace the naming of a wasp to further contour the ongoing nature of data formations in natural history.

Shortly after arriving at the museum, I joined the Hymenoptera (ants, bees, wasps, and sawflies) department of the museum to learn more about how 'biodiversity' is constructed and negotiated in collections and in the practices of scientists working there. It is the largest of the collections in the museum, comprising an estimated 2.2 million specimens. Like the Lepidoptera collection, it is spread across two floors and held in some of the most spectacular, custom-built wooden cabinets. A musty and slightly sweet smell pervades the space. Clusters of modern metal collection storage and the distribution of small QR (quick response) codes on labels, drawers, and cabinets, however, signal a transitional moment. It is here that the first batch of mass digitization of specimens has occurred (more on this below). At the time, I found myself examining a set of shiny wasps, each no bigger than about one centimetre. Thin metal pins had been driven through their bodies, which were stuck to the foam lining of a small white box, a so-called system box or unit pinning tray. Such boxes populate collections in many different sizes and function as a key device for moving, ordering, and protecting specimens (Nadim 2020). Stacks of empty boxes wait in different corners of the collection to be filled, signalling the expected arrival of more specimens. The wasps had recently arrived from northern Thailand, where they had been collected as part of a biodiversity survey carried out by the University of Kentucky. This is not an unusual arrangement, I soon learned. Biodiversity surveys continue the traditional expedition in seeking to establish knowledge of species occurrence in specific areas, mobilizing actors and institutions world-wide while continually filling the empty boxes and jars in museum collections.

Once killed and collected, the animals are roughly sorted in the field before being sent on to respective taxonomic experts in institutions such as natural history museums or universities for further identification. The curator of the Hymenoptera collection invited me to participate in the taxonomic work describing, identifying, and, possibly, naming the wasp, should it turn out to be new to science. A Ph.D. student with whom I was to collaborate had worked on a preliminary survey of the regional wasp diversity and had already narrowed their taxonomic rank to the genus of Alysson, part of the Bembicinae tribe, a group of solitary and predatory wasps that excavate their nests in soil, digging shallow tunnels in which to lay their eggs. The question for us now was whether our wasps constituted new species within this genus, and so we began gathering published descriptions for all known Alysson wasps based on the ultimate reference list, the *Catalogue of Sphecidae*. This was created and is still maintained by Dr Wojciech J. Pulawski, emeritus curator of entomology at the California Academy

of Sciences (the natural history museum in San Francisco). Available online as a pdf document since September 2003, the list remains updated to this day, representing a catalogue of all known Sphecidae wasps and an indispensable instrument for knowing wasps. The oldest description we worked with dated back to 1852 and was published in Latin in *Analecta ad Entomographiam*, a 200-plus-page compilation of entomological descriptions covering the insects of the Russian Empire. In contrast, the most recent description was a 1987 paper from the journal *Acta Entomologica Sinica* written in Chinese. Parsing the materials required interpretation and conjecture – often in discussions with the Ph.D. student and the curator – in order to make them congruent and serviceable for comparisons with the wasps from Northern Thailand. But published descriptions lack comparability: there are no enduring standards for describing morphological characteristics, and there is no agreement on the selection of characteristics to be described in more detail. We therefore embarked on a second-order collection, obtaining the physical specimens behind the descriptions from museum collections in Thailand, the United States, Austria, and the Netherlands. Our wasps were soon joined with more wasps, and I found myself zooming in on and out of many tiny bodies in order to compare veins on wings, abdominal contours, and their published descriptions. Taxonomic work can be a slow process, taking many years to complete.

Historian of science Lorraine Daston has described nomenclature in natural history as an 'art of transmission' (2004: 157), thus shifting the focus from the stability of names as references to the material and semiotic means by which naming is done in practice. The notion of transmission emphasizes the logistical nature of natural history, which has always been concerned with managing the circulation of specimens and names through territories, collections, and inscriptions such as catalogues. This is a slow process which the biodiversity discovery factory intends to speed up. But given the endurance of names, inscribed on museum labels and in catalogues and databases, the transmission might never be entirely unequivocal.

Digitization: testing connections

At the time that I was working on identifying the wasps, the museum embarked on its first large-scale digitization project as a test run for exploring the feasibility of and sociotechnical requirements for translating object-rich – insects comprise by far the largest collections – museum collections into the digital realm, including data releases to GBIF. The project, entitled '*Erschließung objektreicher Spezialsammlungen*' (Establishing an inventory of species-rich collections, EoS) sought to digitize insect drawers, provide virtual access through federated portals, and create efficient and innovative methods for mass digitization of collections, particularly of small, complex objects (Kroupa, Glöckler & Schurian 2015). This spectrum of concerns signals the complexity of digitization, which has become a core operational activity in museums but which remains, nevertheless, under-theorized (Geismar 2013; 2018). By the time I joined the EoS team, the digitization was in full flow. The work took place in a re-purposed office tucked away inside the butterfly collection. Like much of the museum's backstage area, the space is an assemblage of cupboards, shelving systems, and other storage and work furniture spanning three centuries. The central digitization device – a SatScan[3] scanner – rested atop an old wooden pedestal desk. I worked with two digitization assistants to retrieve insect drawers from collections and wheel them to the digitization room using a small trolley. After having moved one of the drawers to my workstation, I carefully opened it by removing the lid. Next, using pinning forceps,

I made sure that all labels were uniformly arranged, orderly pointing in the same direction, because the digital image would capture the entire drawer and its contents, the animal bodies, pins, and labels. Each drawer, species group within the drawer, and type specimen was issued with a QR code that pointed to a unique resource identifier (an alphanumerical code) to retain a connection between specimens, drawers, and digitized objects (more on this below). I then moved the drawer into the scanner and, via a desktop computer attached to it, initiated the scan. A computer screen would gradually reveal the digitized insects as the camera took picture after picture. While I watched the animals appear on the screen, I could see the piles of insect drawers mounting as my colleagues brought in more and more from the museum's collections. That it should be insects that were the first to be subjected to mass digitization betrays a central aspect of the logic of datafication. Insects are, as Raffles writes, 'without number and without end' (2010: 201), signalling the scope of data to be endlessly generated, combined, and multiplied. Like insects, data is conventionally imagined en masse – little is said about an individual insect outside taxonomic work proper as there is little concern over one singular data point in debates around the effects of datafication. In using insects, digitization 'means to ensure their number for [itself]', to appropriate Elias Canetti's (1981 [1960]: 110) reflection on human-insect relations (see Hayden, this volume). Digitizing insects thus spectacularly performs the power of data to go big.

I re-encountered the workflows and the SatScan six years later when the digitization ensemble had indeed been scaled up and moved into the public exhibition area of the museum. The workflows that had been carefully prototyped and tested in the insect digitization now serve as the baseline for the next step in the digitization process, which aims to comprehensively digitize all of the museum's collection in full view of museum visitors. For this, large parts of the Hymenoptera collection were moved into the public exhibition right next to the central dinosaur hall. Beautiful old wooden cabinets filled with bees, wasps, ants, and sawflies bracket one end of the public digitization hall. The other end is furnished with new uniform metal cabinets that await the insects once they have been digitized.

Digitization here is performed as a multidimensional modernizing process. Between past and future, a number of digitizing workstations form an ensemble of humans and machines, surrounded by specimens, labels, pins, papers, and tools. While some stations are dedicated to macroscopic digitization of individual specimens, the mass digitization of insect drawers remains the core process. Once a drawer has been scanned using the SatScan, it is moved to a new station which consists of a custom-built 'digitization street', an ensemble of equipment wound along a small, semi-circular conveyor belt. A digitization assistant removes specimens from the drawer and carefully detaches the labels from the pins. Stacks of QR codes sit on the desk integrated into the station as well as on almost every desk in the digitization hall. They are printed on small strips of rectangular paper and feature a machine-readable label in the form of black squares arranged in a quadratic grid on a white background, the name of the museum, and a unique identifier in the form of a URL (which can be accessed when the label is scanned with a camera or other imaging device). The assistant pins the now label-less insect onto a Styrofoam tray, places a QR code next to it, and neatly arranges the labels. Once the tray is placed on the conveyor belt, it is taken on a quarter-revolution around the semi-circular tracks into a set-up featuring a camera. There it stops for the camera to take three or four images of the specimen and its label(s). Once this is completed, the conveyor belt resumes, taking with it the tray bearing insect, labels, and QR code. At the

end of the conveyor belt, new unit pinning boxes await the objects. Again, QR codes are issued for each box before it is moved into a new drawer that, once more, is issued with a QR code. Finally, these drawers are placed in one of the new collection cupboards.

A key practical and conceptual problem amidst the proliferations of digitization pertains to relations and the question of how things are connected and, importantly, stay connected. The final EoS report stresses the importance of the QR code for the 'provenance of the connection between the physical specimens and the digital representations' (Kroupa *et al*. 2015: 9). The challenge here is not only to connect physical specimens with their paper records but also to trace object references across historical documents and documentation such as catalogues, inventories, and accession books, which are mostly kept in the museum's historical archive and thus do not mirror the taxonomic order of the collections. MacKinney (2019) notes a veritable breakdown in record-keeping beginning in the 1820s-1830s when imperial networks brought in more and more specimens. Rather than noting each incoming individual specimen, museum staff had switched to note shipment numbers instead, thus moving to a transactional order that made numbers and counts the central data element. The digitization compels the museum to revisit, perhaps even confront, these histories (e.g. Heumann, Stöcker, Tamborini & Vennen 2018) as it looks to reconstruct the trajectories of specimens in order to ensure their function as reference.

Next to the QR code, another type of connection is mobilized in the digitization. In a system tray sitting next to the operator of the 'digitization street' rests a single bee. It remains untouched as I watch the operator run the machine, and after encountering it once more as I return to the digitization hall, I enquire about its fate. It is, I learn, the 'test bee', an orphaned specimen from the collection whose place and origin have been lost. It now serves to calibrate the machine when this is switched on. The test bee thus acts as the sole vanguard for the entire class of insects as it is transformed into millions of digital traces.

Conclusion

Data formations have always been a part of natural history (Müller-Wille 2017). Identifying, describing, and naming species and transforming specimens into inscriptions depend on human-data mediations that are often not readily discernible as such, including the careful handling of specimen labels or the painstaking comparison of descriptions and insect bodies. Data formations thus contour the contingent material-discursive constellation of sites, people, and materials that shape bodies and environments through data. Their labours endure beyond the moments and materials of inscription. The scaling up of efforts to discover species, which involve the use of novel molecular technologies, the digitization of museum collections, as well as the transformation of collections into globally accessible databases, adds new elements to these data formations and brings to the fore, with great force, the need for creating and maintaining connections between all elements, old and new. Specimens in the collection require valid and unambiguous names as well as resource identifiers so that they can be mobilized as sources for genetic material which can then circulate in the form of barcodes across databases. Species names on labels and photographs, in pdfs, databases, and descriptions, need to be disambiguated and made to refer to one localized type specimen. Collection objects, biotic and abiotic (drawers, cupboards), are issued with QR codes that need to point to unique resource identifiers, which in turn should connect related objects across ontological domains (wasp and drawer). The

datafication of nature thus proliferates the specification of entities as well as patterns of relations. Asdal suggests that 'the size of nature is not simply there to begin with, but made by way of instruments and connections' (2020: 338). Following on from this, I would suggest that datafication extends the category of nature to encompass colonial histories, institutional cultures, logistics, and technologies, among many other things. So, while scientists might regard the biodiversity crisis also as an information crisis (e.g. Blagoderov, Kitching, Livermore, Simonsen & Smith 2012; Costello, May & Stork 2013) – not having enough data on the world's species occurrences and trends – it might more accurately be framed as a crisis of definition, as Escobar (1999) has argued. To him, the rise of the term 'biodiversity', which I would contend is not coincidental with the datafication of nature, requires an understanding of nature that both is historicized and takes into consideration its economic, social, and political relations.

To conclude, I wish to suggest three ways in which anthropological engagements with scale might contribute to figuring more clearly the stakes in debates around data, big and small. My first point concerns the observation that it is not just that the big contains the small but that the small contains the big (Strathern 1995). Where agency is distributed across human and nonhuman actors, scale becomes an empirical question tied to specific contexts: a butterfly becomes 'large' once folded into a global biodiversity data infrastructure where it can multiply across screens everywhere, while species, the hegemonic ordering logic of life on earth, become a matter of localized documentary traces. Similarly, the world-wide digitization of natural history collections is done through local, often prototypical, human-data mediations embedded in historically specific contexts, guided by idiosyncratic workflows. Importantly, the domaining and magnifying that occur through scaling resist any easy equation of local/diverse/concrete and global/homogeneous/abstract. This does not, of course, mean that such workflows and digitizations are fundamentally tainted or unserviceable. Instead, recognition of the conservation or displacement of heterogeneity as it travels from situated practice to globally accessible data infrastructures (and back) draws attention to the politics of scale. Much consideration has gone into the conventions of taxonomic nomenclature, and debates about biodiversity data integration have given rise to efforts seeking to formalize and standardize data practices. Yet the 'values, norms, interests, and working conditions' (Simons, Lis & Lippert 2014: 636) that are guiding scale production and mobilization, while immanent in, for example, the choice of 'relevant' data points, remain bracketed out of these debates.

My second point considers the issue of context. Scale is a product of contexts, or, as Woolgar and Neyland put it, context has a 'scalar quality' in always being bigger than that which it is said to contain (2013: 108). As a central sense-making device, it has come to dominate much critical engagement with the datafication of life (and lives), also because it facilitates an easy enculturation and socialization of data. The presumed separation of object and background, text and context also plays a role in the construction of biodiversity data and its processes of decontextualization and recontextualization (Leonelli 2016). Here debates as to the proper level of metadata, data about data, for the specimen, its digital representation, and its digitized version are guided not just by arguments of efficiency (in the form of, for example, minimum data requirements) but also by the organisms themselves, the epistemic communities that have formed around them, and the communities of expected and unexpected data users (Kirksey 2015). Some traces of the context of production are retained through the metadata that accompanies data, such as location, collection method, or date;

others are shed – or 'purified' – in order to allow data mobilization and application across many different contexts. Yet the notion of data formation makes evident that any a priori distancing of data and context is anything but self-evident. Data in the context of data formations appears as an 'uncooperative figure' (Ballestero 2019: 20) as it does not easily separate from its background. The taxonomic work of making species (data) shows the difficulty of rendering object-environment distinctions and of drawing boundaries. Similarly, digitization and, more generally, the datafication of nature are enacted in multiple contexts (natural history, museums, global data infrastructures), which make them never only ever part of just one programme, logic, or culture. Again, as with scale, the continual figure-ground reversals compel a sensibility towards the politics of what we might call 'multimodal contexting': the enrolment of materials, actors, networks, and histories in making and appealing to certain contexts and not others.

Lastly, I want to suggest that multisited, interdisciplinary ethnographic engagements can not only furnish more precise understandings of the nature and politics of data and digitization but also offer instructive insights for designing data models, workflows, and data infrastructures. Scholars of infrastructure studies have already proposed ways of collaborating with scientists on ensuring sustainable and usable systems (Edwards, Bowker, Jackson & Williams 2009). At the same time, anthropologists are building their own data infrastructures and inform the development of metadata standards for cultural and social science research (Crowder, Fortun, Besara & Poirie 2020). But there remains more and ongoing work to be done on developing analytical tools to think about the nature, function, and effects of data and digitization. While the dichotomy analogue/digital might suggest digitization to be a straightforward process from one to the other, attention to situated practices of human-data mediations betrays the complexity of digitization as well as the instability of the dichotomy. The material and the digital are intertwined aspects of complex processes and phenomena and not binary opposites (Sumartojo, Pink, Lupton & Heyes le Bond 2016). Similarly, data contains and is contained by manifold sociocultural practices. With the notion of 'data formation', I have tried to convey the sociocultural and historical specificities and contingencies within data.

NOTES

I would like to thank my colleagues at the Museum für Naturkunde Berlin for sharing their insights and curiosity. I would also like to thank Hagit Keysar, Felipe Mammoli, Sarah Blacker, Ingmar Lippert, and Filippo Bertoni for inventive discussions of parts of this essay. Lastly, a special thanks to the fabulous Editors of this special issue for their kind exactitude and exceptional support.

Open Access funding enabled and organized by Projekt DEAL.

[1] GBIF is an expansive resource providing over 1.6 billion species occurrence records: that is, data about the presence of species in specific locations across the globe. Many funding streams stipulate a default data release to GBIF and their datasets feed into high-level policy documents such as Intergovernmental Panel on Climate Change (IPCC) reports (e.g. Warren, Price, Graham, Forstenhaeusler & VanDerWal 2018). It is freely available online through a web browser (at gbif.org) and offers a prominent search bar where users can input a taxon name and retrieve a list (and map) of locations for the species. GBIF data is provided by collections and institutions worldwide, including the Museum für Naturkunde Berlin.

[2] All natural specimens taken from the German colonies were initially sent to the Museum für Naturkunde Berlin, which had been designated the central collection point through a resolution of the Reichstag.

[3] The SatScan was developed by the (now defunct) UK-based company Smartdrive Limited. It resembles a black box measuring about 120 × 70 × 70 cm fitted with a camera suspended from the inside ceiling and running on precision rails allowing it to move in two dimensions to capture images of the object placed

underneath it. The camera takes 256 slightly overlapping images that, in a second step, are stitched together to form a contiguous high-resolution image of the object placed underneath it, in this case insect drawers.

REFERENCES

Asdal, K. 2020. Is ANT equally good in dealing with local, national and global natures? In *The Routledge companion to actor-network theory* (eds) I. Farias, C. Roberts & A. Blok, 337-44. London: Routledge.

Ballestero, A. 2019. The underground as infrastructure? Water, figure/ground reversals, and dissolution in Sardinal. In *Infrastructure, environment, and life in the Anthropocene* (ed) K. Hetherington, 17-44. Durham, N.C.: Duke University Press.

Blagoderov, V., I. Kitching, L. Livermore, T. Simonsen & V. Smith 2012. No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys* **209**, 133-46.

boyd, d. & K. Crawford 2012. Critical questions for big data. *Information, Communication & Society* **15**, 662-79.

Canetti, E. 1981 [1960]. *Crowds and power* (trans. C. Stewart). New York: Continuum.

Costello, M.J., R.M. May & N.E. Stork 2013. Can we name earth's species before they go extinct? *Science* **339**, 413-16.

Crowder, J.W., M. Fortun, R. Besara & L. Poirie (eds) 2020. *Anthropological data in the digital age: new possibilities – new challenges*. Cham: Springer.

Daston, L. 2004. Type specimens and scientific memory. *Critical Inquiry* **31**, 153-82.

Edwards, P., G. Bowker, S. Jackson & R. Williams 2009. Introduction: An agenda for infrastructure studies. *Journal of the Association for Information Systems* **10**, 364-74.

Escobar, A. 1999. After nature: steps to an antiessentialist political ecology. *Current Anthropology* **40**, 1-30.

Franklin, S. 2003. Re-thinking nature-culture: anthropology and the new genetics. *Anthropological Theory* **3**, 65-85.

Geismar, H. 2013. Defining the digital. *Museum Anthropology Review* **7**, 254-63.

——— 2018. *Museum Object Lessons for the Digital Age*. London: UCL Press.

Heumann, I., H. Stöcker, M. Tamborini & M. Vennen (eds) 2018. *Dinosaurierfragmente: Zur Geschichte der Tendaguru-Expedition und Ihrer Objekte, 1906-2018*. Göttingen: Wallstein.

Kirksey, E. 2015. Species: a praxiographic study. *Journal of the Royal Anthropological Institute* (N.S.) **21**, 758-80.

Kitchin, R. 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society* **1**: 1 (available online: *http://journals.sagepub.com/doi/abs/10.1177/2053951714528481*, accessed 15 January 2021).

Knorr Cetina, K. 1999. *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

Kroupa, A., F. Glöckler & B. Schurian 2015. Effiziente Arbeitsabläufe und innovative Methoden zur Erschließung und dauerhaften Verfügbarmachung objektreicher Spezialsammlungen am Beispiel der entomologischen Sammlung des Museum für Naturkunde Berlin. Abschlussbericht. Berlin: Museum für Naturkunde Berlin (available online: *https://www.museumfuernaturkunde.berlin/sites/default/files/abschlussbericht_eos_mfn.pdf*, accessed 15 January 2021).

Latour, B. 1987. *Science in action: how to follow scientists and engineers through society*. Cambridge, Mass.: Harvard University Press.

——— 1999. *Pandora's hope: essays on the reality of science studies*. Cambridge, Mass.: Harvard University Press.

——— & S. Woolgar 1986. *Laboratory life: the construction of scientific facts*. Princeton: University Press.

Leonelli, S. 2016. *Data-centric biology: a philosophical study*. Chicago: University Press.

MacKinney, A.G. 2019. Nature's registry: documenting natural historical collection and trade in Prussia, 1770–1850. Ph.D. thesis, Humboldt-Universität zu Berlin.

Müller-Wille, S. 2017. Names and numbers: 'data' in classical natural history, 1758-1859. *Osiris* **32**, 109-28.

Nadim, T. 2020. System box (tray) with wasp. In *Boxes: a field guide* (eds) S. Bauer, M. Schlünder & M. Rentetzi, 109-23. Manchester: Mattering Press.

Nghonda, J. & S. Zacharie 2007. Colonial cartography as the diplomatic tool in the territorial formation of Kamerun (1884-1916). In *Proceedings of the 23th International Cartographic Conference*, 4-10. Moscow.

Raffles, H. 2010. *Insectopedia*. New York: Pantheon.

Ruckenstein, M. & N.D. Schüll 2017. The datafication of health. *Annual Review of Anthropology* **46**, 261-78.

Schnee, H. (ed.) 1920. *Deutsches Kolonial-Lexikon*, vol. **1**. Leipzig: Quelle & Meyer.

Simons, A., A. Lis & I. Lippert 2014. The political duality of scale-making in environmental markets. *Environmental Politics* **23**, 632-49.
Stoler, A.L. 2008. Imperial debris: reflections on ruins and ruination. *Cultural Anthropology* **23**, 191-219.
Strathern, M. 1995. *The relation: issues in complexity and scale*. Cambridge: Prickly Pear Press.
——— 2004. *Partial connections*. Walnut Creek, Calif.: AltaMira Press.
Subramaniam, B. 2014. *Ghost stories for Darwin: the science of variation and the politics of diversity*. Urbana: University of Illinois Press.
Sumartojo, S., S. Pink, D. Lupton & C. Heyes le Bond 2016. The affective intensities of datafied space. *Emotion, Space and Society* **21**: November, 33-40.
Turnhout, E., E. De Lijster & K. Neves 2014. Measurementality in biodiversity governance: knowledge, transparency, and the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). *Environment and Planning A* **46**, 581-97.
van Dijck, J. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* **12**, 197-208.
Warren, R., J. Price, E. Graham, N. Forstenhaeusler & J. VanDerWal 2018. The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C. *Science* **360**, 791-5.
Woolgar, S. & D. Neyland 2013. *Mundane governance: ontology and accountability*. Oxford: University Press.

La donnéification de la nature : formations de données et nouvelles échelles en histoire naturelle

Résumé

Cet article s'intéresse à la disparition de certaines échelles et connexions tandis que d'autres font leur apparition, dans un contexte où l'histoire naturelle adopte des infrastructures de données digitales. Sur la base de travaux en cours au Musée d'histoire naturelle de Berlin, l'autrice suit la piste des relations entre les spécimens d'insectes et leurs informations écologiques matérielles et numériques. L'autrice applique la notion de « référence circulante » de Latour pour suivre les spécimens d'insectes au fil des taxonomies, des bases de données et des dispositifs de numérisation. En se concentrant sur les médiations humain-données dans les pratiques du musée en matière de commande, de description et de distribution des spécimens, l'article montre comment la « donnéification » de la nature rend présents des contextes conventionnellement dissociés, parmi lesquels le colonialisme allemand. En proposant le concept de formation de données, il suggère que les ethnographes ont une contribution forte à apporter en abordant les aléas et spécificités socioculturelles et historiques que renferment les données.