DNA barcoding and taxonomy: dark taxa and dark texts

Roderic Page

University of Glasgow

November 3, 2015

0.1 Summary
Both classical taxonomy and DNA barcoding are engaged in the task of digitising the living world. Much of the taxonomic literature remains undigitised. The rise of open access publishing this century, and the freeing of older literature from the shackles of copyright has greatly increased the online availability of taxonomic descriptions, but much of the literature of the mid- to late twentieth century remains offline (“dark texts”). DNA barcoding is generating a wealth of computable data that in many ways is much easier to work with than classical taxonomic descriptions, but many of the sequences are not identified to species level. These “dark taxa” hamper the classical method of integrating biodiversity data using shared taxonomic names. Voucher specimens are a potential common currency of both the taxonomic literature and sequence databases, and could be used to help link names, literature, and sequences. An obstacle to this approach is the lack of stable, resolvable specimen identifiers. The paper concludes with an appeal for a global “digital dashboard” to assess the extent to which biodiversity data is available online.

Keywords: DNA barcoding, taxonomy, dark taxa, dark texts, digitisation

1 Introduction
As with many fields, digitisation is having huge impact on the study of biodiversity. Museums and herbaria are engaged with turning physical, analogue specimens into digital objects, whether these are strings of A’s, G’s, C’s and T’s from DNA sequencing machines, or pixels obtained from a digital camera. Libraries and commercial publishers are converting physical books and articles into images, which are then converted into strings of letters using optical character recognition (OCR). Despite the sometimes contentious relationship between morphological and molecular taxonomy, there are striking parallels between the
formation of DNA sequence databases in the twentieth century and the rise of natural history museums in the preceding centuries [38, 39].

Viewed in this way, both classical taxonomy and genomics are in the business of digitising life. Some of the challenges faced are similar, for example algorithms developed for pairwise sequence alignment have applications in extracting articles from OCR text [27]. But in other respects the two fields are very different. Sequence data is approximately doubling every 18 months [21], whereas the number of new taxa described each year has remained essentially constant since the 1980s (see below). A challenge for sequence databases is how to handle exponential growth of data; for taxonomy the challenge is often how to make a dent in the vast number of objects that don’t have a digital representation [7]. This paper explores some of these issues, focusing on taxonomy and DNA barcoding.

2 Taxonomy

Among the many challenges faced by taxonomy is the difficulty of determining the size of the task it faces. Estimates of the number of species on Earth are uncertain and inconsistent, and show no signs of converging [12]. Some estimates based on models of taxonomic effort suggest that two-thirds of all species have already been described [13]. Analyses that use the number of authors per species description as a proxy for effort [19] ignore the global trend for an increasing number of authors per paper [2], and assume that the effort required per species description has remained constant over time. An alternative interpretation is that the quality of taxonomic description is increasing over time [36], reflecting both increased thoroughness and the availability of new technologies [37, 5].

![Figure 1: Numbers of new animal names published each year for animals as a whole, and various taxonomic groups. Data from ION and BioNames.](https://example.com/figure1.png)

Rather than try and estimate an unknown (the number of species remaining...
to be described), here I focus on the current state of taxonomic knowledge. Given that we lack a comprehensive, global index of all species descriptions, discovering what we know about what we know is not entirely straightforward. For zoology the nearest we have is the Index of Organism Names (ION, http://www.organismnames.com), which is based on Zoological Record. Figure 1 shows the numbers of new taxonomic names covered by the International Code on Zoological Nomenclature (animals plus some protozoan groups) that have been described each year based on data from ION, cleaned and augmented in BioNames [28]. These data show an increase in overall numbers over time, with dips around the times of the two World Wars, followed by an essentially constant number each year since the mid-twentieth century. The pattern varies across taxa, some taxa show increasing numbers per year, but other taxonomic groups are essentially static or in decline, even in groups thought to be hyperdiverse such as nematodes [10].

2.1 Digitising the taxonomic literature

The rate of progress in biodiversity research is controlled by two factors, the speed with which we can discover and describe biodiversity, and the speed with which we can communicate that information [32]. Unlike most biological disciplines, the entire corpus of taxonomic literature since the mid 18th century remains a vital resource for current day research. In this way taxonomy is similar to the digital humanities, which has not just “big data” but “long data” [4]. Not only is this because of the rules of nomenclature that dictate (with some exceptions) that the name to use for a species is the oldest one published, it reflects the uneven effort devoted to the study of different taxonomic groups [22]. For poorly known groups the bulk of our knowledge of their biology may reside in the primary taxonomic literature.

Digitisation is one step towards making taxonomic information available. Many commercial publishers have, on the face of it, done the taxonomic community a great service by digitising whole back catalogues of relatively obscure journals. However, digitisation is not the same as access, and many commercial publishers keep this scanned literature behind paywalls. In some fields legal issues around access have been side-stepped by constructing a “shadow” dataset that summarises key features of the data while still restricting access to the data itself. For example, by extracting phrases comprising a set of $n$ words ($n$-grams) from Google Books it is possible to create a data set that contains valuable information without exposing the full text [23]. However for taxonomic work, there does not seem to be an obvious way to extract a shadow. Agosti and colleagues [3, 29] have explored ways to extract core facts from the literature and re-purpose these without violating copyright, though how much of their conclusions can be generalised across different national and international legal systems remains untested.

Apart from commercial digitisation of the scientific literature, two other developments are accelerating access to taxonomic information. The first is the rise of open access publishing, notably journals such as ZooKeys that sup-
port sophisticated markup of the text [31]. This is increasing the number of recently-described species that are published in a machine-readable form that can then be subject to further processing [24]. At the same time, the Biodiversity Heritage Library (BHL) [18] has embarked on large-scale digitisation of legacy taxonomic literature. Although initially focusing on out of copyright literature (i.e., pre-1923 in the United States), BHL is increasingly getting permission from copyright holders to scan more recent literature as well. Coupled with tools such as BioStor [27] to locate and extract articles within the scanned volumes BHL is fast becoming the largest available open access archive of biodiversity literature.

To quantify the extent to which the taxonomic literature has been digitised, for each decade I counted the number of publications of new names in animals both with and without a digital identifier (such as a DOI, a PDF, a Handle, or a URL to BioStor). The recent taxonomic literature is mostly digital: for the years 2010-15 60% of publications have a digital identifier, the bulk of these having a DOI. However, prior to the 21st century more publications lack identifiers than have them, with the 1970s being the least digitalised decade (Fig. 2).

2.2 The long tail of taxonomic literature

Another challenge presented by the taxonomic literature is that it is highly decentralised, being spread across numerous journals (Fig. 3). What is striking is the dominance of animal taxonomy by the “megajournal” Zootaxa, and yet this journal has published only 15% of the new names that have been minted since 2000. The taxonomic literature has a very “long tail” of small, often obscure journals that contain a few taxonomic publications. Long tails require significant effort to index [14], although the Zoological Record claims 90% coverage.
of the taxonomic literature [40], in some taxa there may be significantly greater gaps [11]. Conversely, if we set our sites lower, then long tail distributions mean that we can get a substantial fraction of the names from a small number of journals (the “low hanging fruit”). Indeed, the first 20% of the journals in Fig. 3 contain 80% of the names in BioNames that are linked to a publication. Unfortunately, many of these journals are not currently available digitally (Fig. 2).

The picture that emerges from our knowledge of the taxonomic literature is that recent literature is mostly digital, identified with DOIs, and some of it is open access. But much of our fundamental knowledge of the world’s biodiversity, particularly that published in the mid to late 20th century remains digitally inaccessible (Fig. 2). Between the 21st century trend towards digitisation and open access, and the removal of restrictions pre-1923 as copyright expires, lies a great body of 20th century work that will require considerable effort to make available.

3 Genomics

In contrast with taxonomic knowledge, which is widely scattered, most genomic information is highly centralised, being stored in the three components of the International Nucleotide Sequence Database Collaboration (INSDC), namely GenBank, EMBL, and the DDBJ [9]. Taxonomic name “databases” more closely resemble digitised library catalogues, whereas sequence databases contain the actual sequences, which means we can compute over them. For example, a researcher with a new sequence can discover a lot about that sequence by a simple BLAST search [6], whereas a taxonomist armed only with a name will...
struggle to get computable data from the name alone.

Although the bulk of the world’s sequence data is available in the INSDC, this is not the case for DNA barcodes, most of which reside in the Barcode of Life Data System (BOLD) [34, 35]. Since 2009 BOLD has released nearly 2.5 million DNA barcodes, with updates every few months. However, many of these sequences are not currently available in GenBank. To document this I searched for barcodes in GenBank using two criteria. The first searched for sequences that were listed in the Bioproject database [8] under accession PRJNA37833. The second searched for sequences with the keyword “barcode”. For both searches the sequences were grouped by their date of publication (the “pdat” query parameter) that correspond to the intervals between each BOLD data release. Plotting counts of these sequences over the same intervals as the BOLD data releases highlights the limited data sharing between BOLD and INSDC (Fig. 4).

Figure 4: Comparison of numbers of DNA barcodes in each release of public data from BOLD with the numbers of barcode sequences in the INSDC sequence database GenBank. Data on sizes of BOLD data from http://bins.boldsystems.org/index.php/dataloads, data from GenBank obtained via Entrez query interface.

3.1 Dark taxa

As desirable as data sharing is, it is not without complications. In 2011 I coined the phrase “dark taxa” [http://iphylo.blogspot.co.uk/2011/04/dark-taxon-genbank-in-post-taxonomic.html, see also] Parr [2012] to refer to species in GenBank that lacked formal scientific names. Typically they will have a name that comprises a genus name and some combination of letters and number to make the name unique within GenBank (e.g., a specimen code or the first letter of the last name of the researchers that deposited these sequence).

The pattern shown in Fig. 5 likely reflects a combination of processes. If most of the taxa being added to GenBank represent species that have already been described, then the rate at which taxa can be identified (either by taxonomists or by researchers using their outputs, such as keys) is being outstripped by the pace of sequencing. Alternatively, dark taxa may represent unknown species, but we lack taxonomists capable of recognising the taxa as
new (and formally describing them). If taxonomic capacity is a limiting factor then we would expect a gradual decline in percentage of named taxa, which is the background pattern in Fig. 5. The growth of dark taxa might also reflect changing practices of molecular workers, for example in DNA barcoding where large numbers of specimens are sequenced and deposited into GenBank labelled with specimen codes rather than taxonomic names. Indeed, the dramatic increase in the numbers of dark taxa in 2010 is mostly due to sequences from the BOLD project (recognised by the prefix “BOLD”) being added. Even if we allow for the import of unidentified BOLD sequences as a one-off event, at present less than half the newly sequenced invertebrate taxa being added to GenBank have been identified to species level.

![Figure 5: Growth of dark taxa in GenBank for invertebrate sequences. For each year the graph shows the percentage of species-level taxa added during that year (based on the ‘edat’ query field in Entrez Utilities) that do not have formal scientific names.](image)

4 Integrating biodiversity information

Typically integration across biodiversity databases is achieved using taxonomic names [30], but the rise of dark taxa makes this problematic for an increasing fraction of sequence-based data. Even if we have names, these need not always mean the same thing [20]. As an example, Fig 6a shows the distribution of the lizard *Morethia obscura* from the Global Biodiversity Information Facility (GBIF). For comparison, Fig. 6b shows a geophylogeny [26] for DNA barcodes from BOLD for *Morethia obscura* which reveals considerable phylogenetic structure within “Morethia obscura”. Specimens of this species are assigned several distinct Barcode Index Numbers (BINs) [35] implying that “Morethia obscura” comprises more than one species.

Although GBIF and BOLD present rather different views of the “same”
species, there is considerable overlap in the specimens used to construct Figs 6a and 6b. For example, DNA barcode WAMMS012-10 was obtained from specimen WAMR127637, which also occurs in GBIF (as occurrence 691832260. Because the taxonomic concepts in GBIF and BOLD are explicitly defined with respect to sets of specimens we can directly compare them, rather than rely on the possibly erroneous assumption that a given taxonomic name means the same thing in the two databases. Furthermore, as increasing numbers of type specimens are sequenced [15] we can more firmly associate names with sets of specimens, leading to a computable nomenclature where the name we assign to a set of specimens can be determined automatically [33].

Figure 6: Comparison of Morethia obscura in GBIF (top) with DNA barcodes from the same taxon in BOLD (bottom). Note the phylogeographic structure in the sequence data, which is assigned to several different BINs imply the existence of multiple species within “Morethia obscura”.

Integrating databases using specimens is attractive, but not without its own set of problems. The biodiversity informatics community has yet to standardise identifiers for specimens, despite numerous efforts [17], consequently there may be little apparent overlap between specimen identifiers in different databases.
As an example, despite the limited sharing of data between BOLD and GBIF, there are already barcoded specimens in GBIF. To illustrate, consider the DNA barcode GWORH520-09 from sample “BC ZSM Lep 10234”. GBIF doesn’t have this record from BOLD, but it does have the specimen BC ZSM Lep 10234 [provided by the host institution]ebd55b32-cd68-46b4-85e2-105de99fecc8. The DNA barcode from this specimen is also in GenBank, and because that record is georeferenced it has been ingested by GBIF as part of the Geographically tagged INSDC sequences dataset [1]. Hence, GBIF has duplicate records for a barcoded moth, neither provided directly by BOLD (Fig. 7). Merging and de-duplicating specimen-based records is going to be a significant challenge for global aggregators such as GBIF.

Figure 7: Fig. 7 Illustration of multiple records for the same specimen of *Casbia rectaria* in GBIF. The voucher specimen for the DNA barcode is in GBIF (provided by the institution housing the specimen). The COI barcode sequence from this specimen is in both BOLD and GenBank, and because the sequence is georeferenced it is GBIF as part of a dataset of georeferenced sequences from EMBL.

5 Summary

Both taxonomy and barcoding are actively digitising the living world. The description of new animal taxa is essentially proceeding at a constant rate, generating a steadily growing legacy of taxonomic literature into which digitisation has made modest inroads. In contrast, sequence databases as a whole are growing exponentially, although barcode growth is more modest. Nucleotide sequences are “born digital” and readily computable, for example they can be clustered into BINs of similar sequences, or phylogenies of the type shown in Fig. 6. Given the obvious overlap between the goals of classical taxonomy and barcodes, the lack of digital overlap between these two endeavours is disconcerting. Many barcodes lack taxonomic names (“dark taxa”), and much of the primary tax-
onomic literature has not been digitised ("dark texts"). Integrating barcodes and taxonomy at scale is going to be significant challenge, as indeed will be integrating barcodes into mainstream sequence databases. Mapping between databases using taxonomic names seems the obvious approach, but the abundance of dark taxa shows this has not been entirely successful. Alternatives such as integration via specimens show promise, but are hampered by the lack of stable identifiers. If we are to make process the stubborn problem of the lack of unique, persistent identifiers, and cross links between these identifiers needs to be tackled in earnest [25].

As a postscript, in writing this opinion piece, I have had to write custom scripts to query various databases in an ad hoc manner, trying to extract and assemble information that gives insight into the current state of biodiversity digitisation. For these analyses and visualisations to have broader utility it would be desirable to have some way of consistently and automatically doing these analyses, in effect creating a dashboard of digitisation that would enable us to not only see where we are as a field, but also suggest directions in which we could be heading.

6 Acknowledgments

I thank Paul Hebert for the invitation to speak at the 6th International Barcode of Life Conference. Some of the ideas discussed here were first developed on posts on my blog iPhylo and benefited from feedback from people who left comments on those posts.

7 Competing Interests

I have no competing interests.

References


[31] Lyubomir Penev, Donat Agosti, Teodor Georgiev, Terry Catapano, Jeremy Miller, Vladimir Blagoderov, David Roberts, Vincent Smith, Irina Brake,


