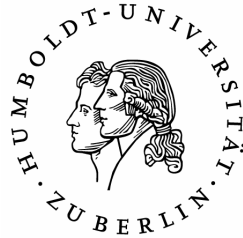


HUMBOLDT-UNIVERSITÄT ZU BERLIN

INSTITUT FÜR BIBLIOTHEKS- UND  
INFORMATIONSWISSENSCHAFT



BERLINER HANDREICHUNGEN  
ZUR BIBLIOTHEKS- UND  
INFORMATIONSWISSENSCHAFT

HEFT 249

**MANAGING MOLECULAR TAXONOMIC DATA**

VON  
JONAS J. ASTRIN



# **MANAGING MOLECULAR TAXONOMIC DATA**

**VON  
JONAS J. ASTRIN**

---

Berliner Handreichungen zur  
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn  
Herausgegeben von  
Konrad Umlauf  
Humboldt-Universität zu Berlin

Heft 249

**Astrin, Jonas**

Managing molecular taxonomic data / von Jonas Astrin. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2009. – 62 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 249)

ISSN 14 38-76 62

Abstract:

Biological taxonomy, the description and re-identification of species, is not able to face the current biodiversity crisis adequately: 70-98% of the several million species on our planet are still undescribed, while extinctions take place every minute. In order to overcome the 'taxonomic impediment', it has recently been proposed to speed up the identification of known species as well as the process of species discovery by using short genetic signature sequences, so-called 'DNA barcodes'. During the last few years, hundreds of thousands of DNA barcodes have been assembled. However, little emphasis has so far been given to theoretical considerations concerning the management of these data.

Here I address questions on how to handle the data needed in molecular taxonomy (e.g. digital preservation, data quality, annotation, database integration). For this purpose, I analyze the NCBI GenBank database and the Barcode of Life Data System (BOLD).

Online-Version:

<http://www.edoc.hu-berlin.de/series/berliner-handreichungen/2009-249>

# Table of contents

<b>1 INTRODUCTION .....</b>	<b>6</b>
1.1 Taxonomy and the current biodiversity crisis.....	6
1.2 The taxonomic impediment.....	6
1.3 DNA taxonomy / DNA barcoding.....	8
1.4 Goals of this study.....	10
<b>2 DISCUSSION .....</b>	<b>11</b>
2.1 Molecular biology databases; GenBank and BOLD.....	11
2.2 Data in molecular taxonomy.....	16
2.3 Annotation standards and information retrieval.....	25
2.4 Identifiers in DNA taxonomy.....	34
2.5 Quality of the information.....	38
2.6 Preservation of molecular taxonomic data.....	42
2.7 Data integration.....	47
<b>3 CONCLUSION .....</b>	<b>54</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>55</b>
<b>REFERENCES.....</b>	<b>55</b>
<b>APPENDIX: INTERNET RESOURCES.....</b>	<b>61</b>

# 1 Introduction

## 1.1 Taxonomy and the current biodiversity crisis

Biological systematics, or biosystematics, deals with the plurality of life and makes it accessible to our understanding. While taxonomy discovers, describes, names and classifies the diversity of organisms, phylogenetics reconstructs their evolutionary relationships. Taxa function as the 'units' of systematics. These are defined categories of organisms at various levels of the 'tree of life', e.g. the species *Homo sapiens* is a taxon, as is the class Mammalia or the kingdom Animalia. Taxonomy assigns specimens to existing taxa (process of determination or re-identification) or to new taxa, which are then delineated and formally described (according to the Linnaean system). As taxonomy provides the most fundamental information on organisms, all life science disciplines depend on it (just to name a few: ecology, evolutionary biology, immunology, pharmacology, agriculture, forestry, etc.).

Taxonomists have so far described approximately 1.7 million species of organisms on our planet (Hawksworth 1995). This took an enormous scientific effort and 250 years to achieve (since the emergence of modern taxonomy and the current nomenclatural system, Linné 1758). Nowadays, roughly 15,000 new species are added per year (Gewin 2002). At first, these seem to be high numbers. However, the situation changes if we consider that an estimated 5 - 80 million species are still unknown to us (May & Lawton 1995) and that we will not be able to catalogue most of this biological diversity if we do not increase our current pace dramatically: today, on an hourly basis, maybe even every minute, biological diversity (biodiversity) is lost due to human ecological impact – much faster than it could possibly form anew. The current extinction rate is unprecedented in biology/paleontology and is rapidly increasing (cf. Pimm et al. 1995).

## 1.2 The taxonomic impediment

To meet the challenge of the above-described 'biodiversity crisis', taxonomy needs to be boosted (cf. May 1990; Mace 2004), facilitating biologically fundamental data that might soon be impossible to gather and giving conservation biology and biodiversity science the necessary scaffold to act upon. However, taxonomy as a discipline has contracted considerably (cf. Ronquist & Gärdenfors 2003) through dwindling numbers of active specialists as well as of funding and popularity (cf. the common analogy to stamp collecting and association with "dusty" collections). The Convention on Biological Diversity ([CBD](#)) that resulted from the Rio Summit (the 1992 United Nations Conference on Environment and Development in Rio de Janeiro) and the scientific community talk about the "taxonomic impediment" in this context. During the last years,

fundamental discussions on taxonomy seized space in journals usually inaccessible to factual taxonomic research due to the publishers' concern with impact factors (most taxonomic literature is rarely cited and read mostly by specialists for the respective group of organisms). Many ways to ameliorate the taxonomic impediment have been proposed and controversially discussed. If taking into account only intrinsically scientific measures (i.e. not considering sociological solutions, e.g. Machlis 1992), these suggestions usually center on one or several of the following aspects:

(i) ***Changing the way in which taxonomic information is managed.*** Today, taxonomy data mostly are scattered through a wide array of sources often troublesome to get to (Polaszek 2005). They are written in a variety of languages and in ignorance of other existing work (Minelli 2003). Suggestions to change this encourage centralization of the taxonomic community's knowledge in order to readily provide up-to-date access to it (Godfray 2002; but see Thiele & Yeates 2002). It has also been proposed to further increase accessibility and dimensions of information via the Internet (e.g. Agosti & Johnson 2002; Wilson 2004).

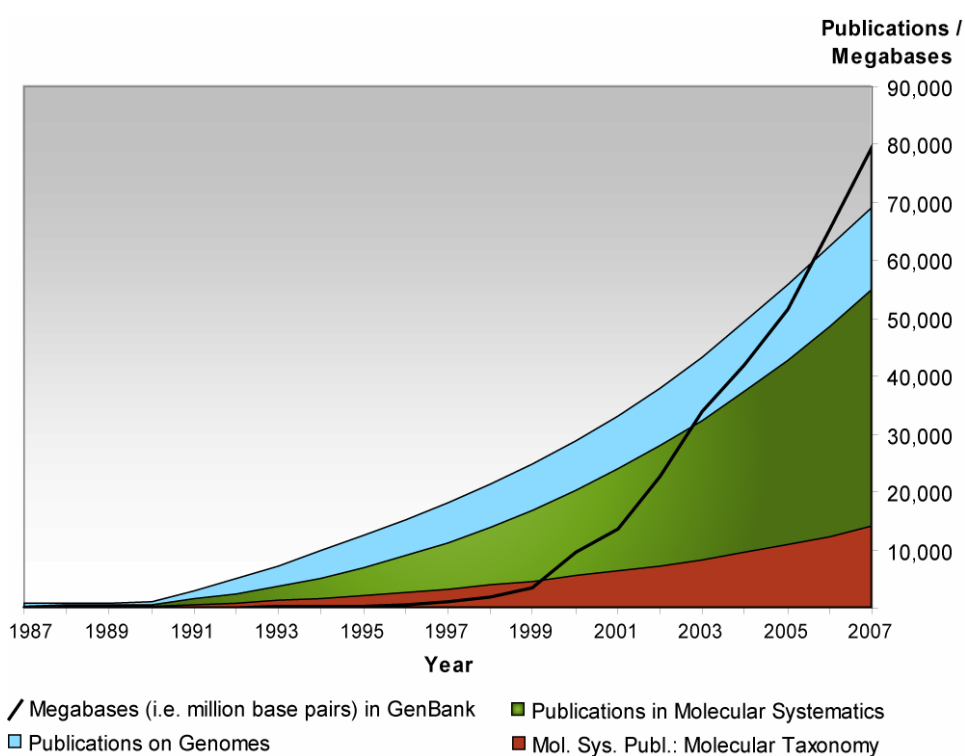
(ii) ***Changing the nomenclatural system.*** Acting as the framework that regulates the process of taxa denomination in systematics, nomenclature has considerable consequences on taxonomy. The most basic principles of nomenclature have to be introduced before some nomenclatural proposals on overcoming the taxonomic impediment can be presented. The principle of priority in the codes of biological nomenclature (simplifying: the valid name for an organism is the one given to it first) serves as criterion when straightening out the frequent cases of synonymy and homonymy. Due to that rule, work creating new or revising existing classification never loses its legal character and hence importance (irrespective of quality). The specimen(s)/culture that serve as evidence for the description of a species must be deposited at a collection (exception: viral taxonomy; bacteria: at two collections). These vouchers are known as 'types'. Different nomenclatural codes apply to each: animals, plants (and fungi), bacteria and viruses. While in plant and animal nomenclature, a published name is already valid if assigned in accordance to the respective codes, bacterial nomenclature dictates a new name be also registered in a specific journal for acceptance. One of the suggestions to improve the current taxonomic situation advocates mandatory species registration also for zoology and botany (Minelli 2003; Polaszek 2005; but see Lughadha 2004). This also connects to the problem outlined in (i). Other incentives suggest resetting priority for groups of organisms while creating a unitary "first web revision" for it as new starting point (Godfray 2002; but see Knapp et al. 2004). Carried by the opinion that biochemical evidence should constitute the substructure of taxonomy (see below), it was also proposed to define new types (neotypes) for known species in cases of inaccessible genetic information from the original type (Tautz et al. 2003; but see Seberg et al. 2003).

(iii) *Change of characters used.* Increasingly, molecular genetic information is gaining support as an ample, independent source of readily quantifiable characters or features (accessible to fast standardized taxonomic analysis and comparable among a wide range of organisms) – for species identification, but also for their description. As was the case during the antecedent discussion of molecular methods in phylogeny, the community is split over just how much room should be given to the application of genetic methods in taxonomy.

### 1.3 DNA taxonomy / DNA barcoding

Molecular genetic techniques in taxonomy were previously applied in eclectic fashion, prevalently to very small or outwardly simple organisms (single-celled organisms, fungi, mosses, algae, roundworms, many parasitic and cryptic taxa, i.e. groups with 'look-alikes'), to specific developmental stages (larvae, spores), parts of organisms (meat or blood samples, root parts, bone fragments) or to mixed samples (environmental or dietary probes). But as deoxyribonucleic acid (DNA) sequencing techniques (that 'read' an organism's genetic, i.e. hereditary information) and computer technology are rapidly improving while at the same time getting more affordable (e.g. Shaffer 2006; Kumar & Dudley 2007; Hine 2008), the method is much more widely applied and the spectrum of organisms treated in molecular taxonomic studies broadens continuously.

Figure 1: Development in number of publications and nucleotide sequence data



The (absolute) number of publications in the field of molecular systematics is depicted in green: it grows almost as fast as the number of publications containing the word "genome". If considering only publications in molecular taxonomy, however, the growth is much smaller. The bold line represents the growth of GenBank (see below): comparing the slopes, it is obvious that molecular studies produce relatively more sequence information as time progresses. Data source: ISI Web of Knowledge - Journal Citation Reports (accessed March 2008) and GenBank Flat File Release 164.0, Feb. 2008.



Figure 1 shows the increase in publications that have resulted from studies in molecular systematics and visualizes technological improvement (mirrored by the mounting number of sequenced DNA base pairs). In addition to this intrinsic (technology-driven) gathering of momentum of molecular methods in biology, further impetus is starting to be – and will be – imparted to DNA-based taxonomy in the light of the above-described taxonomic impediment. Merging at least two [(i) and (iii)] of the above-mentioned incentives, it has recently been proposed repeatedly that genetic methods be applied in taxonomy on a regular basis. Not methodologically new, but unprecedented in the global claim of their approaches, i.e. to systematically cover all – or most – taxa, the topics of 'DNA barcoding', 'DNA taxonomy' (both described below) and related concepts (Blaxter 2003) originally fueled a highly polarized and political discussion. This was maybe due to the lopsided and somewhat oversimplifying form in which they were at first proposed, but was likely also influenced by the apprehension of some 'classical' taxonomists. However, the discussion lost most of its political note with the realization that molecular taxonomy, especially DNA barcoding, does not pose a risk to and partly depends on morphological taxonomy (which studies the organisms' outward appearance), but instead constitutes just an additional tool for the 'integrated' taxonomist and could even hold chances for pure morphological taxonomy (Gregory 2005).

Tautz et al. (2003) present a model of '*DNA taxonomy*' that aims at typifying all species also molecularly, ideally based on sequences of several standard genes. In this model, fragments of genetic sequences are to be used for the definition and re-identification of species and could build a stable link between a species and any type of information (taxonomic, phylogenetic, ecological, etc.) associated with it. As a reference system, they could also help to "guard against duplicate descriptions" (Tautz et al. 2003) thanks to their reproducibility and objectivity (but see Lipscomb et al. 2003). (Note: if written without quotes, I use the term DNA taxonomy not as a concept like Tautz et al. 2003, but as the theory and practice of answering taxonomic questions through DNA sequence analysis, i.e. as narrower term or de-facto synonym to molecular taxonomy.) Another concept, '*DNA barcoding*', is advocated by Hebert et al. (2003a, b) and is already being put into practice, mostly through the Consortium for the Barcode of Life ([CBOL](#)). CBOL encompasses more than 150 member organizations over all continents. Its secretariat is located at the Smithsonian Institution. (The corresponding 'bottom-up' approach is represented by the [Barcode of Life Initiative](#).) Originally proposed only for animals, the barcoding technique was later adapted also to fungi, protists and plants (e.g. Chase et al. 2007). DNA barcoding centers on speeding up species identification through the use of short fragments of molecular sequences. These signature sequences or 'DNA barcodes' (in free analogy to the symbology used for trade items by the Universal Product Code system or the International Article Number) are usually

obtained in a single sequencing run from only one standard gene, but often, the routine application of more than one gene has been advocated. After an adequate number of barcodes have been deposited in a database, these function as profiles when assigning unidentified DNA sequences to known species – or they help in flagging potential new, i.e. undescribed species for further investigation. At least in theory, this constitutes a fast, reliable and standardized way to obtain information on the species of an unidentified organism – or at least on its phylogenetic context (its 'kinship'). Taken together with the aspect that taxonomic specialists could then relegate time-consuming routine identifications to robotic DNA processing and sequencing (and that any non-specialist could perform routine identifications himself/herself on any given group), the compilation of biodiversity inventories could thus be greatly facilitated.

Critical factors for the success of the barcoding initiative are the sufficiency of the sequenced gene(s), effective analytical and database search strategies (Kress & Erickson 2008) and especially, comprehensive barcode 'libraries' (Ekrem et al. 2007). At this point, we are still far away from such comprehensive databases that will have to contain millions of sequences, but sequences are added at considerable speed (cf. Hajibabaei et al. 2005; barcoding records in the BOLD database – see below – currently cover more than half a million specimens, and of these almost 400,000 are associated with a barcode sequence). This is achieved through many small, independent barcoding projects, but also through a rising number of 'big science' enterprises (e.g. [All-Leps Barcode of Life](#), [Fish Barcode of Life](#), [All Birds Barcoding Initiative](#), [Marine Barcode of Life](#), [Polar Barcode of Life](#), [Census of Marine Zooplankton](#), [Sponge Barcoding Project](#), [Canadian Barcode of Life Network](#)).

## **1.4 Goals of this study**

From the above, it can be concluded that molecular information, embedded into an integrative taxonomic framework, is a highly necessary component for tackling the biodiversity crisis – and a useful tool for life sciences in general. Already, a high and quickly growing data volume has been assembled for molecular taxonomy. So far however, little emphasis has been given to theoretical considerations concerning these data and their management.

Which kinds of data do we need in molecular taxonomy? How should we handle these data? How should they relate to other data relevant in taxonomy? I intend to address these questions by discussing selected problems and by analyzing existing nucleotide (i.e. DNA) sequence databases regarding these issues. I will center on the GenBank database and the Barcode of Life Data System, which are discussed in the next chapter.

## 2 Discussion

### 2.1 Molecular biology databases; GenBank and BOLD

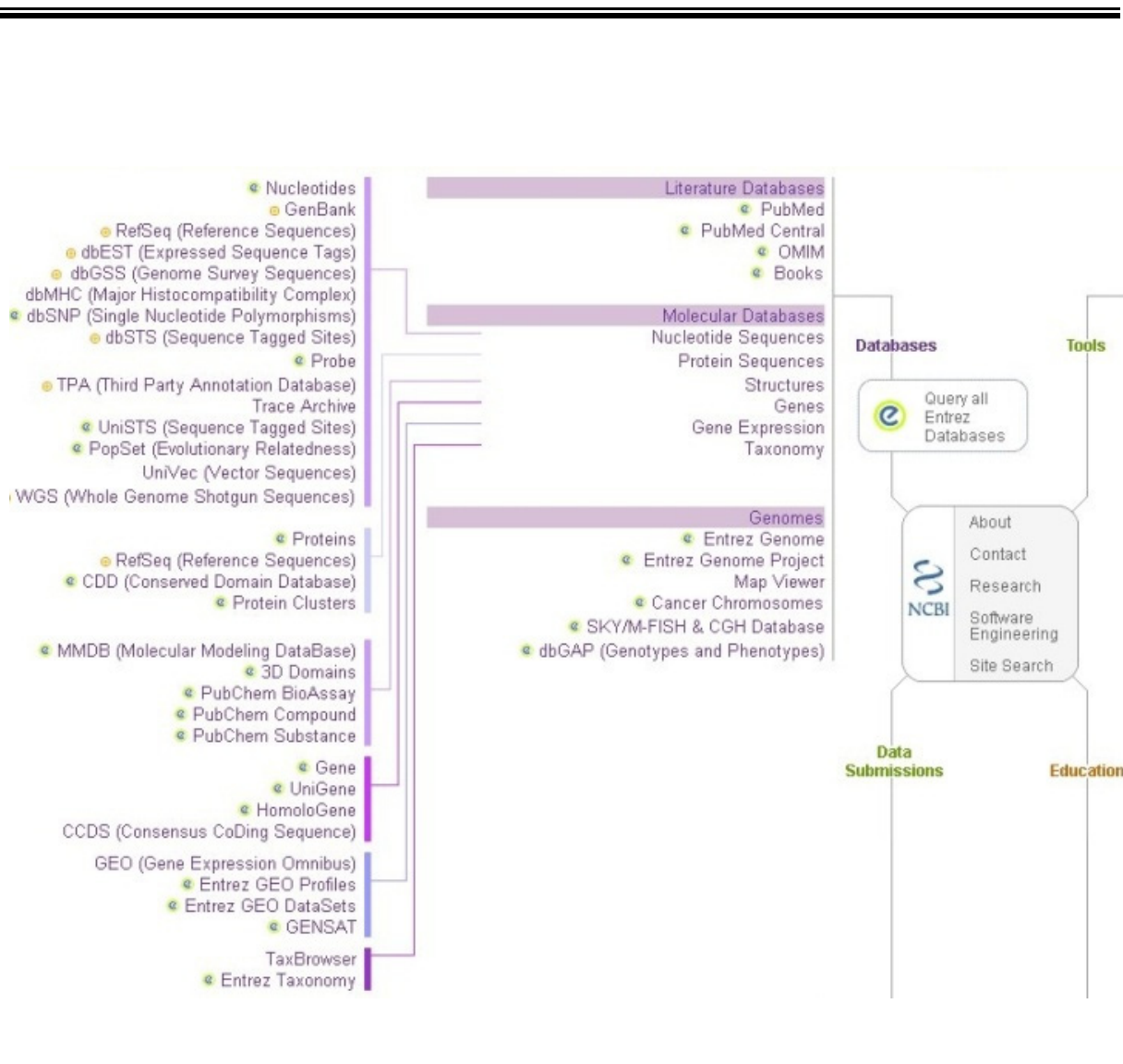
Roughly 20 years after discovery of the DNA overall structure in the 1950s, the first rapid sequencing methods were developed, but the 'data explosion' in molecular biology reached truly industrial dimensions in the 1990s, when sequencing entire genomes (especially the human genome) became a scientific goal and technology was advanced correspondingly. To access the resulting vast amounts of data, at least three elements were necessary: databases, retrieval algorithms and the Internet. I focus on the first one of these three: molecular databases.

The Molecular Biology Database Collection currently lists 1,078 active databases in the field of study (Galperin 2008). The majority of these emerged during the last decade (for the year 1996, Hansen 2004 mentions 60 databases). Most of the 1,078 databases are tailored to a group of organisms, to specific molecules, gene(s), functions or diseases. There exist a plethora of databases on the sequence of molecules (protein or nucleotide), on three-dimensional structure, on enzymes, on molecular nomenclature, on metabolic or cellular signaling pathways, on interactions between molecules, on drug design, on polymorphisms (variations within an organism or a group of organisms), on genomes, genomics or proteomics (comparing entire complements of genetic material or of proteins) and on further aspects. Almost all of these databases connect in some form, directly or indirectly, to the fundamental pivot in molecular biology: the information on nucleotide sequence contained in the macromolecular 'strings' of DNA (or RNA; sequence example: Adenine-Guanine-Cytosine-Adenine-Adenine-Cytosine-etc. or simply AGCAAC, cf. Fig. 4, top). It is due to this fact that one database occupies a central position in molecular biology and related branches: the database compiled by the International Nucleotide Sequence Database Collaboration ([INSDC](#)), which has the objective to collect any publicly known nucleotide sequence from the community and make it freely available. Although this is a single database when considering its raw data content, on an organizational level it is split into three individual repositories that can be considered as reciprocal data mirrors (even though they differ slightly in data format). The INSDC partners are (i) the National Center for Biotechnology Information ([NCBI](#), part of the National Library of Medicine; Bethesda, MD, USA) with the [GenBank](#) database (Benson et al. 2008), (ii) the European Molecular Biology Laboratory ([EMBL](#); Heidelberg, Germany – and in this context especially its outstation at Hinxton, UK) with the EMBL Nucleotide Sequence Database (also: [EMBL-Bank](#); Cochrane et al. 2008) and (iii) the National Institute of Genetics ([NIG](#)) with the DNA Data Bank of Japan ([DDBJ](#); Sugawara et al. 2008). GenBank, EMBL-Bank and DDBJ merge new or updated content daily.

As with most information referred to above, molecular taxonomy data are always connected – at least to some extent – to the INSDC. At present, this dependency is the more significant as current molecular taxonomy draws overwhelmingly on the evidence of DNA sequence information, obtained from individual genes or gene fragments. Thus, it seems appropriate to deal with one of the INSDC databases in this context. Here I focus on GenBank, as NCBI's product is by far the most frequently used of the three INSDC contributors' databases.

One of the earliest bioinformatics projects, GenBank was maintained at Los Alamos National Laboratory (Los Alamos, NM, USA) from its creation in 1983 until the early 1990s,

Figure 2: Map of Entrez data sources



Part of the GenBank sitemap: Entrez sources are shown, while the Tools, Data Submission and Education sections can be accessed at: <http://www.ncbi.nlm.nih.gov/Sitemap/>

when it was transferred to the NCBI (NLM: NCBI Handbook, accessed Feb. 2008). Its original location at Los Alamos with the lab's then immense computer resources shows the strong connection between molecular biology and computers, as pointed out by Smith (2002). GenBank size (in terms of base pairs or nucleotides) has approximately doubled every 1.5 years<sup>1</sup> since 1982 (cf. Fig. 1), when it contained around 680,000 base pairs in a little bit more than 600 sequences. Now it comprises more than 89 billion base pairs in 86 million sequences (NCBI-GenBank Flat File Release 165.0, April 2008; not counting whole genome shotgun sequences). This sequence information is derived from more than 170,000 distinct species (NCBI Taxonomy database, accessed Feb. 2008 using the most conservative query options). Around 200,000 users accessing GenBank make 4 million database queries per day, "making this site second only to the U.S. Internal Revenue Service in amount of use" (Mount 2004). GenBank is accessible to the user through the elaborate metasearch engine and web portal [Entrez](#), which currently connects NCBI's 35 major databases and links to its tools (Wheeler et al. 2008). These encompass bibliographic and databases and a wide range of molecular databases (cf. Fig. 2). One of the possible output forms of a GenBank record is exemplified in Figure 5.

At the beginning of this chapter, some possible contents of molecular biology databases were enumerated. But in addition to content, they can also be classified according to type: GenBank is a primary or 'archival' (e.g. Berendsen 2003) database, as it gathers the raw (experimental) data 'fresh' from the laboratory. Usually, the sequences are submitted to GenBank (or any other repository of the INSDC) by the researchers prior to or during publication process. Another category is constituted by the secondary, 'derived' or 'curated' databases. These obtain their information by computationally and/or manually processing primary databases (e.g. the universal protein resource [UniProt](#)). Databases catering to a specialist community or targeting specific organisms, genes, etc. are called specialized databases (Xiong 2006). Explicit DNA taxonomy databases belong to this last category. Specialized databases can sometimes be a 'hybrid' between a focused secondary and a primary database, when workers submit their data directly to the specialized database. Such a hybrid database is the Barcode of Life Data System ([BOLD](#); Ratnasingham & Hebert 2007).

Representing an instrument of the DNA barcoding movement, BOLD emerged only a few years ago (after a short testing phase of the first version, the second version was launched in 2006). BOLD is so far the only specialized molecular taxonomy database for 'higher organisms'. It

---

<sup>1</sup> This brings into mind "Moore's law" which states that the number of transistors on a computer chip double every two (or, in another opinion, every 1,5) years, but – as alluded to above – sequencing technology also accounts massively for nucleotide sequence database growth, as do general factors connected with sociology of science (cf. de Solla Price 1963).

was developed and is maintained at the Canadian Centre for DNA Barcoding / Biodiversity Institute of Ontario ([CCDB](#); Guelph, Canada). Whereas GenBank deals with all kinds of DNA sequence information, BOLD focuses on DNA barcodes from only one gene (the cytochrome oxidase subunit I or COI gene from the mitochondrion; however, BOLD can also handle a gene from the cell nucleus: the Internal Transcript Spacer and more can be added). At present, BOLD catalogs approximately 400,000 barcode records for almost 40,000 species – most of these animals, but also plants, protists and fungi (bacteria are usually identified through another gene). Roughly a year ago (iBOL web page, accessed April 2008), BOLD delivered around 35,000 query hits per day. Currently, BOLD has somewhat more than 2,400 registered users (S. Ratnasingham, pers. comm.). The aim of BOLD, both database and online workbench, is manifold (but I will focus mostly on the database aspect throughout the discussion). It acts as a repository for DNA barcode data and for the associated information on the source specimen along with collateral data (cf. Fig. 6 for a typical specimen record and Fig. 11 for part of a barcode record). The hereby resulting barcode 'library' is meant to serve public users in identifying their unknown specimens through DNA barcodes (via an '[Identification Engine](#)'). Furthermore, BOLD can be used to manage and analyze data within an ongoing project – e.g. by a network of different labs that simultaneously work on a common, larger project (e.g. the [Fish Barcode of Life Initiative](#), cf. BOLD project [list](#)). The statistics shown in Figure 3 are part of these management functionalities. Finally, BOLD, as an interface to GenBank, aids in the preparation of special 'BARCODE' records for submission to the INSDC's collective database. BARCODE records have to fulfill certain criteria. They especially have to contain more information on the source specimen of the DNA than usually supplied in GenBank (they also have to give account on how the sequence was obtained technically, etc.). The integration of these workbench elements defines the actual character of BOLD and made the BOLD operators change its name from originally the 'Barcode of Life Database' to the current 'Barcode of Life Database System' (or also Systems).

BOLD and GenBank both are realized through a relational database structure. They offer their database content and search functionalities through the Internet (GenBank offered its content on Compact Disc format until 1998). The GenBank database can also be completely downloaded as ASCII text files. GenBank currently has one mirror site (BOLD does not yet, but is planning several, see chapter 2.7). Submissions for GenBank and BOLD can be carried out online or sent in via email (as a spreadsheet in BOLD; GenBank offers a program for download for this purpose). BOLD offers a tutorial (which GenBank does not) and uses an approach that allows the user a more intuitive navigation than in GenBank. This is important as BOLD is meant to also receive users with limited 'molecular literacy', looking for identification of their barcode sequences. However, it often lacks help functions (always provided in GenBank, even though sometimes



hard to find) to explain inherent concepts, methods and terms. A feature that GenBank implements (whereas BOLD does not yet) is the personalization of its web interface ([My NCBI](#) allows permanent sessions, standardization of searches and filters, etc.).

Figure 3: Management component of BOLD for a particular project



A section of a project overview page ("Fishes from South China Sea") from the BOLD Management and Analysis component.

Source:

<http://www.barcodeoflife.org/views/projectlist.php>

Now do we really need both GenBank and BOLD? **Both databases are necessary, as they have different aims and advantages.** The broad archival nature and centralization effect of GenBank is necessary especially for users outside the barcoding or biodiversity community (thematically and

temporally). BOLD is necessary for DNA taxonomy through the specialized access it provides, as it both filters away irrelevant molecular information and bundles information sources for the needed data. Beyond its database functions (i.e. through its project management component) it can furthermore direct research, indicating cases in which further collecting or (re-)analysis is necessary as well as preventing the duplication of results. BOLD contains more information than GenBank (e.g. digital images, cf. Fig. 14) and offers more sophisticated ways to visualize this information (e.g. by mapping coordinates). But most importantly, BOLD keeps different records on sequence data and source specimen (whereas GenBank focuses primarily on the sequence). In taxonomy, it is of paramount importance to keep track of the specimen in order to be able to re-access it and to convey meaning to the molecular results. The 'biphasic record' is BOLD's strength, and it should take care to keep this system up without blurring the borders<sup>2</sup>.

## **2.2 Data in molecular taxonomy**

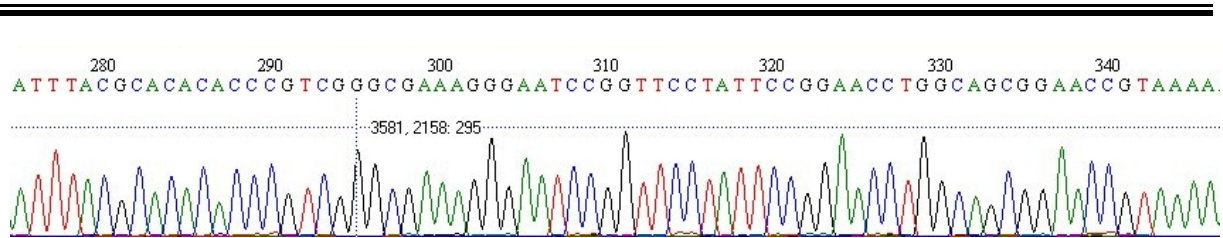
The previous chapter outlined possible recipients for DNA taxonomy data. But what are these data? I define them as those informational units resulting from or necessary for taxonomic investigation. In DNA taxonomy, the *primary* evidence consists of nucleotide sequence *data*. What are these? The automated sequencing machine (the 'sequencer', increasingly often a robot) delivers as output an electropherogram that is obtained by laser scanning the fluorescent-labeled copies of the DNA sample (on sequencing theory, see e.g. Hillis et al. 1996). Such an electropherogram (or sequence trace) features a string of peaks which, based on their color and distinctness, have been associated to nucleotide bases through the computerized base-calling function of the sequencer environment. Figure 4 shows an electropherogram fragment. After sequencing, usually two electropherograms (the two separately sequenced complementary strands of a DNA molecule) are arranged one above the other with matching positions and a consensual sequence (called 'contig' in genetics) is determined. This contig consists of a sequence of a very limited number of ASCII-encoded letters – if no ambiguities exist in the sequence, it is constituted

---

<sup>2</sup> When dealing with repositories, BOLD does not always clearly separate between the repository of the specimen and that of the sequence data. When data is not submitted directly to BOLD but retrieved from GenBank and no specimen repository is specified, BOLD enters 'GenBank' as 'Specimen Depository' (cf. any of the GenBank projects in BOLD's [project list](#)). Further: the corresponding specimen identifiers are assigned the GenBank accession number by BOLD (which is associated primarily with the sequence in GenBank). And finally, the barcode 'counter' on the [initial page](#) mixes up statistics from databases (GenBank) and institutions holding specimens (Canadian barcoding center).



Figure 4: Exemplary section of an electropherogram



Letters denote individual nucleotide bases of the sequence, numbers on top give the position within the sequence (for human readability). Chromatogram peaks correspond to signal type and intensity as detected by the sequencer.

by only four symbols (NC-IUB 1985). With some interest on the topic from the part of molecular biology, there could be ample grounds for controversy, as it is not a straightforward matter which is to be considered the primary data: the original hereditary information as sublimated in the organism's nucleotide sequence, or its transformation into graphic information (trace file), or the transformation into textual information (contig). One could argue that all three are – at least for preservation. But while the original biological evidence and the trace file have to be kept mostly for the purpose of reproducing results, the starting point for analysis is the pure text sequence (usually extracted from the database in special sequence [formats](#), e.g. FASTA; cf. Mount 2004). It is also to this textual sequence information that the metadata are attached to. Pragmatically, the textual information can be seen as the primary evidence.

The data explosion of DNA sequence data in relation to the number of publications (cf. Fig. 1) makes it impossible to conventionally publish these data and in some cases even creates difficulties to link them to publications (cf. Ashburner & Goodman 1997; the latter does not yet apply to taxonomy). If we further consider the pervasive data sharing that substantially aided in the success of molecular biology (Parr & Cummings 2005) and consider also the need for computational analysis of sequence data, we realize that when we deal with DNA sequence data, we will be dealing with digital data stored in databases. Aside from this, DNA sequence data are born-digital data due to the automation of the sequencing process. Thus, the mentality often found in natural science practitioners (among others) that "if it's not on the Internet, it doesn't exist" gains a factual (or should one say 'virtual?') reality in this context.

The *metadata* in DNA taxonomy could be grouped into four different classes.

(1) One class describes the *sequence*, e.g.:

- what is the 'name' of the sequence?
- what is its identifying number/code?

- where is the sequence information stored?
- who submitted the sequence and when?
- how was the sequence obtained?
- which genetic locus (or loci) is it derived from?
- where do I find further information on the specific locus (e.g. identifier for external genetic information)
- does it translate into a product and where does the read start and/or end?
- what function does it have?
- what can be said about the nature and cellular provenience of the macromolecule (DNA/RNA, circular/linear, nuclear/organelle etc.)?
- is there any indicator for the quality of the sequence?

Some sequence metadata can be calculated from the sequence itself, e.g.:

- how long is the sequence?
- what is the amino acid sequence of the protein product (if any)?
- in which proportion do its components (nucleotides) occur?

(2) Further metadata can be provided regarding the *specimen*, e.g.:

- who collected it?
- how, when and where (GIS) was it collected and under which conditions (e.g., association with other organism(s), abiotic factors)?
- where do I find further information on the specific collecting locality (e.g. identifier for external geographic information)
- what was the age, life stage/ecobiomorph and condition of the specimen?
- which sex does the specimen belong to?
- how many years after collection was the DNA isolated from the specimen (e.g. in material taken from natural history collections)
- where is the specimen/tissue voucher archived that served as DNA source?
- what is the voucher's name/identifier?
- where is this isolated DNA archived?
- what is the DNA voucher's name/identifier?
- which tissue type or part of the organism was the DNA extracted from?
- was the DNA isolated from more than one individual (simultaneously)?

(3) Another class of metadata relates to the hypothesis made on *taxonomic placement*:

- what taxon is the sequence assigned to (e.g. species or genus)?
- who made this assignment and when?

- which evidence was the assignment based on (DNA sequence or previous, e.g. morphological or bioacoustic, analysis of the source specimen)?
- where do I find further information on the specific taxon (e.g. identifier for external biodiversity and biodiversity resources information – or if classification has changed)

(4) The last group of metadata are the *bibliographic* data:

- in which study – if any – was the data (or the results obtained from it) originally published? (The publication itself, i.e. its content, should be considered as secondary data.)

More metadata could be added to this list for special cases or as taxonomic methodology changes over time. Of course, not all of the listed metadata always apply to each case, and usually, few sources will annotate – i.e. record metadata – in this detail. In fact, it can be argued that some of the information will be excessive depending on the respective project. For example, it is time- and memory-consuming in a database that features only one or a few gene(s) like BOLD to include too much information on the gene (function, provenience, etc.) in the individual records.

Whereas many of the above metadata categories apply also to other branches of molecular biology, the correctness and depth of annotation concerning the specimen and the systematic placement are of fundamental importance to taxonomy research.

***What data are needed in DNA taxonomy? How are these represented in GenBank and BOLD?*** We have to differentiate between two different usages of DNA sequence data in taxonomy (be it in an inclusively or exclusively molecular approach): the *delineation* (description, synonymization or transfer) of taxa or the *re-identification* of already described taxa. The former, especially if not integrated with other evidence, requires as much information as possible from different sources, i.e. ideally from several specimens and from several genetic loci within each specimen. As can be understood when considering the proportion of yet unknown species (cf. chapter 1.1), the usual scenario of species description is built upon specimens that are new to science and that have never been entered into a database. Nevertheless, there are frequent cases in which an existing classification is revised and where database entries already exist (so that species description can be based on specimens that already 'found their way' into a database). Can BOLD records be used in tackling such cases? BOLD specifically aims to include sequences of more than one specimen per species. However, the amount of genetic information per specimen is very limited, as currently only part of a single gene is analyzed. This does not warrant the necessary number of independently evolving characters for investigations concerning species status. As all characters in BOLD are obtained from the same locus, the possibility to corroborate or refute a species hypothesis by multiple independent genetic markers does not exist. Accordingly, formal taxonomic changes (e.g. species description or species synonymization) based on such weak evidence would be extremely unsatisfying and in danger of being revised

soon. What about GenBank? GenBank sometimes includes sequences of several genes that originate from the same specimen (although it can be difficult to logically connect these sequences, as information on the source organism is often absent, or is incorporated either into the 'specimen\_voucher' or the 'isolate' field through the respective submitters). It also often includes several specimens per species (although this is usually the case in those few well-studied model species that are ubiquitous in molecular biology research and can mostly be considered as taxonomically 'stable'). However, GenBank does not usually feature several specimens per species that have each several homologous DNA sequences in the database. Even if we found an instance where both requirements coincided, specimen metadata would likely be insufficient from the perspective of the nomenclatorial codes (e.g. voucher not traceable), prohibiting any formal act (if not dealing with viruses). Thus, we could not, even after obtaining obvious results, join (synonymize) species 'a' with species 'b' since we cannot name specimens to vouch for this hypothesis. For the same reason, we could not describe one or more subsets of species 'c' as the new species 'd'.

From this, I conclude that *neither BOLD nor GenBank lend themselves to primary research in alpha taxonomy* (i.e. species-level classification). BOLD and BOLD-derived records in GenBank could, however, be used in a system that integrates additional evidence if some minor metadata issues were solved (see below).

In species identification, for which barcoding is meant to deliver one of the possible 'master keys', a very short stretch of DNA sequence often suffices to assign an unknown to a known species (even shorter than 'barcodes', which are usually slightly more than 600 base pairs in length; cf. Hajibabaei et al. 2005; Min & Hickey 2007). Many different genes could be applied to many different groups, but some degree of standardization seems reasonable, as the success of a comprehensive DNA identification system depends on extensive coverage of organisms (Ekrem et al. 2007). Thus, a 'horizontal genomics' approach between a multitude of labs and research projects that covers as much of the tree of life as possible for a few or at least one predefined gene fragment seems to be a sensible development. BOLD takes a step into that direction by amassing sequences from the gene COI. GenBank, through its archival nature, also offers the relevant primary data (although it has to be filtered for the appropriate content first). However, the most

Figure 5: Metadata in GenBank

```


LOCUS      EU525415                900 bp    DNA        linear    VRT 24-MAR-2008
DEFINITION Gelochelidon nilotica voucher AJB 6157 cytochrome oxidase subunit 1
            (COI) gene, partial cds; mitochondrial.
ACCESSION  EU525415
VERSION    EU525415.1  GI:169882570
KEYWORDS   BARCODE.
SOURCE     mitochondrion Gelochelidon nilotica
  ORGANISM Gelochelidon nilotica
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria;
            Aves; Neognathae; Charadriiformes; Laridae; Gelochelidon.
REFERENCE  1 (bases 1 to 900)
  AUTHORS   Tavares,E.S. and Baker,A.J.
  TITLE     Single mitochondrial gene barcodes reliably identify sister-species
            in diverse clades of birds
  JOURNAL   (er) BMC Evol. Biol. 8 (1), 81 (2008)
  PUBMED    18328107
REFERENCE  2 (bases 1 to 900)
  AUTHORS   Tavares,E.S. and Baker,A.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-FEB-2008) Natural History, Royal Ontario Museum, 100
            Queen's Park, Toronto, Ontario M5S 2C6, Canada
FEATURES   Location/Qualifiers
     source          1..900
                   /organism="Gelochelidon nilotica"
                   /organelle="mitochondrion"
                   /mol_type="genomic DNA"
                   /specimen_voucher="AJB 6157"
                   /db_xref="taxon:126865"
                   /country="Australia: Western Australia"
                   /lat_lon="19.424 S 120.46 E"
                   /collection_date="02-Apr-1996"
                   /collected_by="A. J. Baker"
                   /PCR_primers="fwd_name: LTyr, fwd_seq:
                   tgtaaaaaggwctacagcctaacgc, rev_name: AspH, rev_seq:
                   ctatgtaatgggtttactaac"
     gene          <1..>900
                   /gene="COI"
     CDS          <1..>900
                   /gene="COI"
                   /codon_start=1
                   /transl_table=2
                   /product="cytochrome oxidase subunit 1"
                   /protein_id="ACA97302.1"
                   /db_xref="GI:169882571"
                   /translation="VTF INRWL FSTNHKDIGTLYLIFGAWAGMVG TALSLLIRAE LGQ
                   PGTLLGRDQIVMVIWVTHHAEVMTTFEMVMPDITGCGFMHIVPI.MTC&PDM&FPRMNNMS

```

Screenshot of a GenBank flat file; red box indicates the section with the 'source modifiers' reside. Note: GenBank hides empty fields. Source: <http://www.ncbi.nlm.nih.gov/Genbank/>

relevant question in this context is whether the metadata supplied in the databases suffice. In species identification, after retrieving an identical or close 'hit' from the reference database that includes only the analyzed gene, the next most relevant information is what the reference sequence stands for, where it comes from and how reliable it is. For the 'normal' GenBank entries, in which data on the source specimen are not regularly available, the answer seems simple: the metadata do not suffice. However, from 2006 onwards (one study already in 2005: Greenstone et al. 2005), GenBank, EMBL and DDBJ routinely offer a specific type of record for molecular identification purposes (using the COI gene) that is labeled 'BARCODE' in the keyword field.

Figure 6: Metadata in BOLD


Perichares of ACG [MACGP]

**Specimen Identifiers**

<b>Sample ID :</b>	05-SRNP-30958	<b>Museum ID :</b>	05-SRNP-30958
<b>Isolate / Field Num:</b>	05-SRNP-30958	<b>Collection Code :</b>	
<b>Donated By :</b>		<b>Deposited In :</b>	University of Pennsylvania

**Taxonomy**


<b>Identifier :</b>	Daniel H. Janzen
<b>phylum :</b>	Arthropoda
<b>class :</b>	Insecta
<b>order :</b>	Lepidoptera
<b>family :</b>	Hesperiidae
<b>subfamily :</b>	Hesperiinae
<b>genus :</b>	Perichares
<b>species :</b>	<i>Perichares prestoeaphaga</i>

**Specimen Details**

<b>Voucher Type :</b>	
<b>Tissue Type :</b>	
<b>Extra Info :</b>	
<b>Sex :</b>	M
<b>Reproduction :</b>	S
<b>Life Stage :</b>	A
<b>Note :</b>	Areaceae

**Collection Data**

<b>Collectors :</b>	Manuel Rios
<b>Date Collected :</b>	27-Mar-2005
<b>Country :</b>	Costa Rica
<b>State/Province :</b>	Guanacaste
<b>Region/County :</b>	Area de Conservacion Guanacaste
<b>Sector :</b>	Sector Pitilla
<b>Exact Site :</b>	Sendero Evangelista
<b>Latitude :</b>	10.987
<b>Longitude :</b>	-85.421
<b>Coord. Source :</b>	
<b>Elevation/Depth :</b>	660



Screenshot of a BOLD specimen page. Source:

<http://www.barcodinglife.org/views/projectlist.php>



The introduction of the BARCODE category was based on a [proposal](#) of CBOL (Hanner 2005). Entries featuring this reserved keyword (in April 2008: roughly 14,000 records) fulfill additional requirements, e.g. they must include information on vouchers and collecting locality (at least country; encouraged: latitude/longitude through a newly introduced field). In the following, I will discuss some of the more relevant metadata and their inclusion or exclusion from the databases. GenBank BARCODE entries (cf. Fig. 5) and BOLD records (cf. Fig. 6) offer the following metadata: what species does the reference sequence belong to? GenBank links the species information to a unique identifier resolved in its taxonomy browser, whereas BOLD presently links to its taxonomy database using only the scientific name (the Linnaean binomen). This might create problems with changing classifications over the years (Kennedy et al. 2005). Further: who was the species identity of the reference sequence determined by? (Both databases offer the identifier field, although in GenBank it is often left empty if sequences are not submitted via BOLD.) When was the reference specimen identified and through which evidence (supported by neither)? Although the evidence used might be too tedious to list, at least a date would be helpful, as classifications change over time (e.g. a species might be split into two, one conserving the old name; now was the specimen determined before or after?) and the process of identification often shifts to different characters (e.g. it is realized that the femoral hair tufts that have been used, say, until 2004, to key out a number of hypothetical species are in fact misleading). If the barcoding system is meant to last, these data should definitely be included.

The so far analyzed metadata were on taxonomic placement (cf. metadata class "(3)" in the above list). Sequence metadata (1) are usually not especially relevant in this context, as the genetic source is standardized. However, a sequence identification number needs to be provided for stability ("Barcode ID" in BOLD, accession number in GenBank) as well as some measure of sequence quality: both BOLD and GenBank encourage submission of trace files, which in BOLD are directly linked to the sequence and measured/classified by an internal algorithm (not illustrated; it uses PHRED base calling scores as described by Ewing & Green 1998, Ewing et al. 1998)<sup>3</sup>. Further sequence metadata, mandatory for barcodes in both databases, are constituted by the information on the DNA 'primers' used to technically obtain the reference sequence (these enable the specific polymerase chain reaction, or PCR). Bibliographic data (4), given in both databases where applicable, will be discussed during the next chapter. A highly important group of metadata fields center on the source specimen (2): who collected it and when (both databases)? Where was it collected? Exact georeferenced coordinates have to be supplied in BOLD, and at least the country needs to be given in GenBank. Coordinates are the most fundamental collecting

---

<sup>3</sup> Other than through metadata, the BOLD identification engine derives a quality estimate through the recovery of at least three matching reference records that fulfill barcode criteria (Ratnasingham & Hebert 2007).

information and should be made mandatory in GenBank BARCODE records, too. The GenBank field 'lat-lon' should be extended to include information on elevation (although provided in BOLD, the field "Elevation/Depth" is often left empty, not only in Figs. 8, 14). Elevation can directly give information on relevant ecological parameters. In addition to this aspect, if querying external databases, the elevation warrants an accuracy check for the location data. Further: where is the specimen archived (only BOLD)? Where is its isolated DNA archived (provided by neither)? Which is the voucher identification number (both databases)? Which is the DNA voucher number (neither of the two databases)? For traceability, at least the specimen voucher should be identified together with its repository. For scientific reproducibility and out of considerations to spare the specimen (which could possibly, some years later, be one of very few left on the planet), the DNA voucher, which is usually deposited by researchers along with the specimen, should be identified in analogy to the way in which the specimen voucher is identified. Further useful information can in some cases be derived from the association of the organism with another species (possibility of DNA contamination or key identification parameters for specialized organisms, e.g. was the beetle found on shrub species 'a' or 'b?'). Neither database currently offers such an option, although resourceful users in BOLD have used the 'Note' field (cf. Fig. 6; nevertheless, information deposited in such a way is hard to search for, as a computer will not be able to put it into a semantic context, especially since the field is grouped with the specimen data and not the collection data and will usually contain other information). Other ecological parameters could also be introduced (maybe the users should be encouraged during submission to report taxon-specific necessities for new fields).

Summing up, I come to the conclusion that ***GenBank BARCODE records and especially BOLD entries are suited to most current species identification demands***. Nevertheless, both (especially GenBank) need to adapt and include more data fields. More standardization (especially BOLD) is needed as well as a way to obtain necessary or mandatory information from the submitters (especially for those fields qualifying for barcode status in BOLD; the status requirements currently do not seem to be strictly enforced, as the incomplete record exemplified in Fig. 8 has been transferred to GenBank as BARCODE entry). Crucial in this aspect would be the mutual adjustment of criteria to suffice barcode status in BOLD (Ratnasingham & Hebert 2007) and GenBank (Barcode Submission Tool, accessed April 2008). These coincide in many aspects, but still differ significantly.

In order to meet future developments in taxonomy, BOLD will have to prepare for two new kinds of data and for considerable challenges: (i) huge amounts of 'environmental' barcodes, i.e. sequence data generated by mixing up assemblages of organisms, e.g. from soil samples, water or sediment (cf. Liu et al. 2007), and sequencing them together. Since its inception three



years ago (Margulies et al. 2005), the pyrosequencing technique has been applied to genome sequencing and microbial diversity studies. Its application to biodiversity of 'higher' organisms is a matter of (very short) time. In a few hours, a current pyrosequencer performs several hundred thousands of DNA sequence reads in a massively parallel manner. With such a technique applied to environmental samples, we can gain stunningly high numbers of DNA barcodes, but without any other information on specimens than the collective geo-spatial and ecological information for the whole environmental collecting lot. BOLD will have to deal with an industrial level of identifications if querying the database with these sequences. And what is more, it will have to deal (at least at first) with many more queries that have no known or even similar counterpart in the database than known. This might be the moment for BOLD (less of a problem for regular GenBank records) to decide whether all 'barcodes' in its database will have to be connected with a interim or regular scientific name or if a new class of derived, environmental barcodes that is connected only to an identifier will be introduced (cf. LSID in chapter 2.4). Most certainly, these data will – in some form – be gathered in a centralized database and funding procured for this purpose. (ii) The second challenge awaiting BOLD, although in a more distant future, will be the standardized sequencing of entire genomes for a wide range of organisms. If disregarding all storage (trace files) and annotation issues, still the bioinformatics involved in simultaneously evaluating and comparing genomes among that many organisms will be highly complex. Will genomes, if integrated into BOLD, move the focus of the platform from pure species identification to also include functions for species description (and for this purpose, add more morphological content)? This seems probable.

### ***2.3 Annotation standards and information retrieval***

With a fast-growing number of sequences in DNA taxonomy, the way of storing and interconnecting data for retrieval becomes an essential task for the future of the discipline, not only for species identification, but also for alpha taxonomy and for many adjoining scientific branches. Analysis and modeling of all available data can create emergent knowledge. The chances for effective knowledge discovery are good due to the convergence of all relevant DNA taxonomy data in the digital medium, i.e. due to their common digital nature, which makes them easily searchable and combinable with other data sources. However, the effectiveness of such knowledge discovery is partly hampered by not adopting data standards necessary for cross-discipline data transfer and by annotations that are not meaningful to computers. The former chapter already went into the different types of metadata in DNA taxonomy. Here, I want to consider their standardization (regarding document structure and content).

In molecular biology, metadata – the annotations – are usually included in the same document as the sequence (i.e. the primary data). The resulting document can be described in various formats. Several structured formats are in use (cf. Mount 2004). GenBank originally used the flat file format shown in Figure 5 to store its data, but to increase search efficiency now only uses it as the standard output file. The format used internally by GenBank is the Abstract Syntax Notation number One ([ASN.1](#)), which was developed by the French computer industry in another context and is a joint standard of the International Organization for Standardization and International Telecommunication Union (ITU-T Rec. X.680 (2002) & ISO/IEC 8824-1:2002). ASN.1 provides tagged fields for information similar to Extensible Markup Language ([XML](#)) or the Genetic Data Environment (Smith et al. 1994) format. It uses a tree-like arrangement (flatter in hierarchy than regular XML) to structure the data and thus enhances computer access. It is not easy, but possible to read ASN.1-formatted data by eye (cf. ASN.1 example in Fig. 7).

Figure 7: Example of a GenBank record in ASN.1

```

        title "ORM1*Q0koeln=H.sapiens gene A for alpha1-acid glycoprotein
exon 4" } } ,
    reftype no-target } ,
    update-date
    std {
        year 2000 ,
        month 4 ,
        day 14 } ,
    source {
        org {
            taxname "Homo sapiens" ,
            common "human" ,
            db {
                {
                    db "taxon" ,
                    tag
                    id 9606 } } ,
            orgname {
                name
                binomial {
                    genus "Homo" ,
                    species "sapiens" } ,
            mod {
                {
                    subtype isolate ,
                    subname "German" } } ,
            lineage "Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo" ,

```

ASN.1 output  
for GenBank  
accession  
AB014887.1  
Source:  
[http://www.nc  
bi.nlm.nih.go  
v/Genbank/](http://www.ncbi.nlm.nih.gov/Genbank/)

Little information has been published as to how BOLD stores its data. BOLD will soon use the relational database management system [DB2](#) offered by the company IBM (S. Ratnasingham, pers. comm.). DB2 used to keep its data in tablespaces inside the system, but from DB2 version 9 onwards, it can handle XML data natively, conserving the inherent logical structure. Currently,

BOLD uses the collaboratively developed database management system [PostgreSQL](#), which is also able to handle XML natively. Both BOLD and GenBank encode their various metadata in clearly defined and searchable fields. So the problem is not so much how to standardize the structure of the document, but how to adapt its structure to (automatically) communicate with other databases, for example in biodiversity. GenBank could improve the communication with other, non-molecular databases by adopting biological data standards for the most relevant BARCODE metadata categories and by implementing these within its ASN.1 framework. Information on the natural history collection holding a specimen and information on the natural occurrence could be conveyed by the XML schema '[Darwin Core](#)', which was developed by the Biodiversity Information Standards group ([TDWG](#); formerly: Taxonomic Database Working Group). The TDWG is an international not-for-profit collaboration concerned mainly with the exchange of biodiversity data. Its secretariat is located in Hobart, Australia. Ratnasingham & Hebert (2007) mention that the BOLD specimen data meet the Darwin Core standard. A more elaborate XML schema that equally focuses on specimen metadata was developed by the TDWG together with a task group of the Committee on Data for Science and Technology ([CODATA](#)): the 'Access to Biological Collections Data' ([ABCD](#)) schema. Another XML data standard applicable to specimen metadata is the Geography Markup Language ([GML](#)) by the Open Geospatial Consortium ([OGC](#)), although a less complex subset of GML would suffice for the needs of DNA taxonomy to encode collecting location data. Metadata describing which taxon a sample belongs to are more ambiguous through the difficulties created by alternative classification schemes (cf. Beach et al. 1993; Kennedy et al. 2005; Graham & Kennedy 2007) and the variability and thus elusiveness of the 'species' (Krichevsky 2005). GenBank and BOLD could benefit from integrating the TDWG standard 'Taxonomic Concept Transfer Schema' ([TCS](#)) which is designed to address data exchange needs between various data systems with varying classifications. Its advantage is that it does not depend on a single official name (plus synonyms) but can couple a set of unique identifiers (see chapter 2.4) that are associated with different taxonomic concepts. Thus, TCS accommodates the aspect that species represent nothing more (and nothing less!) than explanatory hypotheses (cf. Fitzhugh 2006). Bibliographic data: BOLD and GenBank would improve by formalizing their unstructured bibliographic references using at least metadata element sets like the [Dublin Core](#) standard maintained by the Dublin Core Metadata Initiative ([DCMI](#)). However, the biggest problem associated with bibliographic data is that, as authors usually submit their data to GenBank prior to publication, many forget to update their information after publication. Hence, bibliographic data are often almost not present and there would not be much to describe with Dublin Core or other elements (GenBank might consider programming 'bots' that search the Internet for a publication matching the data; these bots could establish

contact with submitters via the email address indicated during submission and in case of no answer relay to the GenBank staff). If not structuring bibliographic data in the DNA sequence database itself, at least an identifier should be provided as link to a (structured) bibliographic database. GenBank implements this in at least some records through referring to [PubMed](#) Unique Identifiers in a reserved field ('PubMed'). BOLD neither structures its references (publications appear in erratic format, sometimes like in Fig. 8, sometimes not), nor links to a bibliographic database. Instead, it sometimes links – in a scientific sense laudably, but legally problematic – to a PDF-file of the publication stored on its server. Sometimes it also links through to the article at the web page of the respective publishing house (which, by the way, could be done more elegantly and sustainably by programming a simple query to the DOI-resolving proxy in the same way BOLD queries GenBank).

The previous section considered the standardized, *machine readable* structure of data with the aim to facilitate data exchange<sup>4</sup>. Another issue concerns semantics: how can we convey meaning to the data when it is read by computers (rendering data *machine understandable*). This, as well, is a matter of annotation standards: which data – and how much – is entered where into the database? Therefore, *submission support and monitoring functions* are required (offered in GenBank and BOLD, but too parsimonious) that explain the different database fields and modifiers and – if standardizeable – specify which controlled vocabulary or diction can be employed (e.g. BOLD sex determination "M" for male, or GenBank date in the form "01-Oct-2006"). Where non-standardized (or non-standardizeable), such a database function should be 'intelligent' to some degree so that it could warn the submitters if they were about to submit information that might also fit into another field (cf. the 'isolate'/'voucher\_number' example above; but there is room for more mix-up in GenBank, e.g. the [source modifiers](#) 'isolation\_source', 'environmental\_sample', 'metagenomic', 'specific\_host', 'lab-host', 'type', 'group', 'subtype', 'subgroup', 'variety', 'biovar', 'chemovar', 'pathovar', 'ecotype', 'cultivar', 'breed'). To date, no such measure has been adopted by GenBank. In BOLD, the danger of losing information through entering it in different fields is reduced due to its lower number of overall data categories – consequence of a smaller and homogeneous user community (however, as demonstrated in the last chapter, the number of data categories in BOLD is still insufficient). Such a submission

---

<sup>4</sup> What will not be considered here in depth (since it is only indirectly relevant for the database) is the standardization needed in the process of exchange itself, i.e. for protocols: usually in molecular biology (and increasingly in biodiversity) these are [SOAP](#) (W3C) or [CORBA](#)-based (OMG). In biodiversity, initially [Z39.50](#) (ANSI/NISO) was borrowed from the library sector (Soberon 1999), but was later replaced by the 'home-made' [DiGIR](#) protocol, which handles Darwin Core, and its modification [BioCASE](#), which handles ABCD (cf. Canhos et al. 2004). [TAPIR](#) (TDWG) integrates DiGIR and BioCASE.


monitoring system should also keep track of which relevant fields are filled out (especially for granting barcode status, see above). The major factor that stands against complete manual annotation is time. Thereby, records like the one obtained from BOLD and shown in Figure 8 emerge. A data harvesting function (cf. Swaminathan et al. 2005), integrated into a global laboratory information management system (LIMS), could automatically fill in much of the relevant data (e.g. in Fig. 8, the name collector and the identifier could be automatically

Figure 8: incomplete BOLD record with barcode status

Specimen Identifiers			
Sample ID :	USNM 627657	Museum ID :	627657
Field Num :		Collection Code :	USNM
Deposited In :	Smithsonian Institution		
Publication :	Comprehensive DNA barcode coverage of North American birds, K. C. R. Kerr, M. Y. Stoeckle, C. J. Dove, L. A. Weigt, C. M. Frances, P. D. N. Hebert, doi: 10.1111/j.1471-8286.2006.01670.x (pdf)		
Donated By :			

Collection Data	
Collectors :	
Date Collected :	01-Jul-1994
Country :	Iceland
State/Province :	Keflavik
Region/County :	
Sector :	Naval Air Station
Exact Site :	
Latitude :	64.0158
Longitude :	-22.3426
Coord. Source :	
Elevation/Depth :	

Specimen Details	
Voucher Type :	
Tissue Type :	
Extra Info :	B09873
Sex :	U
Reproduction :	
Life Stage :	

Taxonomy	
Identifier :	
phylum :	Chordata
class :	Aves
order :	Passeriformes
family :	Motacillidae
genus :	Motacilla
species :	Motacilla alba

Barcode Identifiers			
Barcode ID :	KKBNA204-05	Sample ID :	USNM 627657
Gene :	COX1	Translation Matrix :	Vertebrate Mitochondrial
		Last Updated :	2005-08-26
		GenBank Accession :	DQ433815

Screenshot from a combined specimen plus, only partly shown, barcode record as obtained when following the link to BOLD in GenBank. Order rearranged.

added while they submit their specimen data into the system; however, this would not work if using insufficiently labeled natural history collections holdings). Especially BOLD, which offers a simple LIMS to improve communication between different labs, is not far away from achieving this. The fundamental step is to really make the LIMS comprehensive and global, meaning that it

must be the primary recipient for any information contributed by any of the project members or collaborators. Considering the rapidly increasing throughput of sequencing facilities and the prospected (already beginning) massive gathering of samples from 'biodiversity observatories', 'exploratories' or 'all-taxa biodiversity inventories' for specific sites (Gewin 2002; EDIT 2007), no way seems to lead around sophisticated, flexible LIMS (cf. Brazma et al. 2006). Databases in DNA taxonomy, highly dependent on ample (and specialized) metadata, should closely monitor and contribute to their development. But let us come back to the issue of *controlled vocabularies*. Usually, the spectrum of the respective vocabularies for the different fields is (or where none exists: could be) relatively small and of no hierarchical complexity. Therefore, no state-of-the-art thesaurus or ontology seems necessary. But there exist two data categories that could benefit from adopting one: geographic names and taxon names. Developing thesauri or ontologies requires immense effort and know-how (cf. Schulze-Kremer 2002; Bard & Rhee 2004; Barker & Wu 2005), but luckily, there already exist instances that DNA taxonomy can draw upon. Standardized geographic information could be derived from the Thesaurus of Geographic Names ([TGN](#)) developed and maintained by the [Getty Trust](#) (neither the GEOnet Names Server – [GNS](#) – by the U.S. National Geospatial Intelligence Agency and the U.S. Board on Geographic Names nor the various gazetteers available offer the same degree of granularity and/or clear hierarchical structure delivered by the TGN). For taxon names, useful sources exist that can be integrated: the [Thesaurus](#) used in the [Zoological Record](#) by Thomson Scientific, the fungal resources [MycoBank](#) and [Index Fungorum](#), or the International Plant Names Index ([IPNI](#)). [ZooBank](#) will be a valuable source for zoological names, especially when submission to the animal name registry of the International Commission on Zoological Nomenclature ([ICZN](#)) is made a mandatory component for species description with the next edition of the International Code of Zoological Nomenclature (Polaszek 2005). Bacteria are not considered in barcoding, but are formally [registered](#), too. At present, the most comprehensive index of species names is the 'Catalogue of Life' ([CoL](#), see chapter 2.7). NCBI curates its own [taxonomy database](#), on which the BOLD [Taxonomy Browser](#) is partly based. Although not a recognized taxonomic authority (Barker & Wu 2005), NCBI's taxonomy is a quasi-standard in molecular biology. Maybe this is because it can be downloaded, because it allows to list synonyms and misspellings, and because it features the lineage of an organism (cf. Fig. 7, bottom). Often, this will be enough in molecular biology research. More than 50 new species are added daily to NCBI's taxonomy database (Wheeler et al. 2008). However, its classification is curated in irregular intervals based on submissions and on ad-hoc decisions that are derived from the scientific literature or sporadic expert statements. As good as such a classification can get, the resulting documents are not suited for data exchange with sources that do not draw on the NCBI taxonomy database as well (cf. Page 2006). Beyond taxon names and



geography: what else is there to standardize? Although the names for the few genes that are currently employed in barcoding do not need a structured controlled vocabulary, gene names should nevertheless be standardized. Unfortunately, this is not easy in molecular biology (Smith 2002), where a controlled vocabulary spanning all archival bioinformatics resources has been demanded repeatedly (e.g. Berendsen 2003). The collaboratively maintained [Gene Ontology](#) is a start, but does not focus on gene names, either (it targets gene product attributes). Usually in molecular biology, when genes are associated with an ID, this is according to gene order, i.e. its occurrence within the genome. Thus, they are referenced within and not among organisms<sup>5</sup>. Being faced with the difficulty that no unique naming system for genes is implemented throughout molecular biology, it seems highly advisable to name the sequence during submission according to an agreed-on standard in the barcoding community. The now typical barcoding gene, COI or if spelled out the 'cytochrome c oxidase subunit I' gene could also be submitted as 'subunit I of cytochrome c oxidase' or, as frequently done, omitting the 'c': 'cytochrome oxidase subunit I'. It can also be abbreviated in different forms: COI, CO1 or cox1. This ambiguity in gene denomination can cause problems. E.g., I argue that the number of COI sequences in GenBank is higher than determined by BOLD, as the BOLD search algorithm does not account for all possible entry names: when looking for COI sequences of the weevil *Listronotus bonariensis*, BOLD cited 7 hits from GenBank, while a separate search (using the species name and "subunit I") in GenBank retrieved 13 entries – of which only 7 feature the abbreviated form 'COI'<sup>6</sup>.

***BOLD and especially GenBank make little use of the many available means to standardize structure and content of the data.***

---

<sup>5</sup> To illustrate this: the COI gene in the baker's yeast has the ID number Q0045; in *Schizosaccharomyces pombe*, another yeast, it is SPMIT.01. A huge effort is put into conveying unique names to genes by the [Nomenclature Committee](#) of the Human Genome Organisation ([HUGO](#)), but is again species-specific (i.e. for *Homo sapiens*). At this moment, more research in 'horizontal genomics' seems necessary in order to single out constants and variables before associating genes with unique identifiers (e.g. the life science identifiers discussed in chapter 2.4; cf. Brazma et al. 2006 on the topic). NCBI's [Entrez Gene](#) database can be used to search effectively for specific genes, but does not solve the standardization issue in naming sequences.

<sup>6</sup> Afterwards, I was informed that BOLD searches are not carried out using the name of the whole sequence entry, but instead are based on the gene name field: "COX1[*gene name*] OR COI[*gene name*] OR CO1[*gene name*]", pers. comm., S. Ratnasingham. If adding to the search "OR cytochrome oxidase subunit i[*Protein Name*] OR cytochrome oxidase subunit 1[*Protein Name*] OR cytochrome c oxidase subunit i[*Protein Name*] OR cytochrome oxidase subunit 1[*Protein Name*]", BOLD could considerably enlarge its records. I carried out such a search and – even after removing all hits for "genome" – found that the number of retrieved sequences more than doubled the figure given by BOLD on its counter.

Figure 9: Entrez Nucleotide query form

The screenshot displays the NCBI Entrez Nucleotide search interface. At the top, the NCBI logo is on the left, and a decorative DNA sequence graphic is on the right. Below the logo, there are tabs for 'All Databases', 'PubMed', 'Nucleotide', 'Protein', and 'Genome'. The 'Nucleotide' tab is active, and the search bar contains the text 'barcode[Keyword]'. To the right of the search bar are buttons for 'Preview', 'Go', and 'Clear'. Below the search bar, there are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The 'Preview/Index' tab is selected.

On the left side, there is a blue sidebar with various navigation links: 'About Entrez', 'Entrez Nucleotide Help | FAQ', 'Entrez Tools', 'Check sequence revision history', 'LinkOut', 'My NCBI (Cubby)', 'Related resources BLAST', 'Reference sequence project', 'Search for Genes', 'Submit to GenBank', and 'Search for full length cDNAs'. The 'Search for Genes' link is highlighted.

The main content area contains instructions for using the search form:

- Enter terms and click Preview to see only the number of search results.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use #search, e.g., #2 AND #3 or click query # for more options.

Below the instructions, there is a 'Search' section with a 'Most Recent Queries' list:

- #72 Search Rattus norvegicus
- #69 Search Wizard of Oz
- #66 Search Latimeria

Underneath is an 'Add Term(s) to Query or View Index:' section with instructions:

- Enter a term in the text box, use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index.

The 'Add Term(s) to Query or View Index:' section includes a text box with 'Keyword' selected in the dropdown menu, and buttons for 'Preview' and 'Index'. Below the text box, there is a list of search fields:

- Accession
- All Fields
- Author
- EC/RN Number
- Feature key
- Filter
- Gene Name
- Genome Project
- Issue
- Journal
- Keyword
- Modification Date
- Organism
- Page Number
- Primary Accession
- Primary Organism
- Properties
- Protein Name
- Publication Date
- SeqID String

At the bottom right, there are links for 'Write to the Help Desk', 'NCBI | NLM | NIH', 'Department of Health & Human Services', 'Privacy Statement', and 'Freedom of Information Act'. The name 'Fertig' is visible at the bottom left of the page.

One of the possible query forms in GenBank (cf. also 'Limits' tab on GenBank web page). The drop-down menu shows some of the fields that can be searched through the mask. Source: <http://www.ncbi.nlm.nih.gov/Genbank/>



If all the above-discussed standardizing elements were in place, biodiversity-oriented information retrieval from molecular sources should be straightforward and unhindered data flow in an integrated biodiversity platform would seem possible (see chapter 2.7). Within molecular biology, many forms of data retrieval are applied (e.g., cf. the NCBI data mining [tools](#)), but in DNA barcoding, the usual searches performed on the primary data, i.e. the sequence itself, will be the sequence similarity comparisons. Thereby, users are able to check the species identity of a (new) query sequence introduced by them. Also, among others, the possibility of contamination can be tested or the genetic distances within and between species can be explored, e.g. highlighting cases in need of further research (be it due to research errors or due to interesting evolutionary contingencies). GenBank employs the Basic Local Alignment Search Tool ([BLAST](#)) and derived algorithms to perform sequence similarity searches (Altschul et al. 1990; Wheeler et al. 2008). As BOLD currently only contains COI records that translate into a protein, it makes use of a Hidden Markov Model (Eddy 1998) for sequence alignment (i.e. for ascertaining evolutionary homologies) before performing a tree-based database search (using the neighbor joining algorithm; Saitou & Nei 1987). Other methods could (and should) be applied, for example self-organizing classifications based on heat-maps (Garrity & Lilburn 2005), character-based systems that go beyond pure distance estimates (cf. Sarkar et al. 2002; Kelly et al. 2007; Rach et al. 2008) or several alignment-independent approaches (e.g. Chu et al. 2006, B. Misof, pers. comm.). In any case, the current single-method approach used by BOLD should be supported by further analyses to assess the consistency of the identification. Besides sequence search functionalities, many other queries can be conceived. It should be possible to access any field of the databases through a query. Many categories, but not all (see Fig. 9), can be searched online in Entrez. Compensating the missing search functions, the database can be downloaded entirely and the search tailored to individual needs. The BOLD database, which is not offered for download, has much narrower search options (Fig. 10), probably due to its homogeneous user community. These concentrate on sequence and specimen identifiers, taxon, geography and sequence length. When one consults the multitude of different metadata discussed in chapter 2.2, of which many if not all need to be searchable, it is obvious that BOLD could do better at search options. Another issue is that while Entrez increases flexibility through Boolean operators and wildcard search, BOLD cannot handle any of these.

***Information retrieval from GenBank is satisfactory for general purposes, but BOLD will have to considerably expand its search options and, if aiming at establishing its role as 'the' identification source in DNA taxonomy, will have to focus on additional sequence retrieval algorithms.***

Figure 10: BOLD query form

**BOLDSYSTEMS** | Management & Analysis

**BOLD Search [NO CODE]**

**BASIC SEARCH :**

**Taxonomy**

Phylum : Phylum

Class : Class

Order : Order

Family : Family

Subfamily : Subfamily

Genus : Genus

Species : Species

**Geography**

Country/FOA : Country

State/Province : Province

Basic Search Cancel

**ADVANCED SEARCH:**

**Taxonomy**

Include :

Exclude :

**Geography - Country/Province**

Include :

Exclude :

**Geography - Region**

Include :

**Sequence Length**

Min : Max :

**Specimen/Sequence**

Sampleid :

Processid :

Include GenBank Data

Single Representative Per Species

Screenshot showing the query form that is offered by the BOLD Management and Analysis component. Source: <http://www.barcodinglife.org/views/projectlist.php>

## 2.4 Identifiers in DNA taxonomy

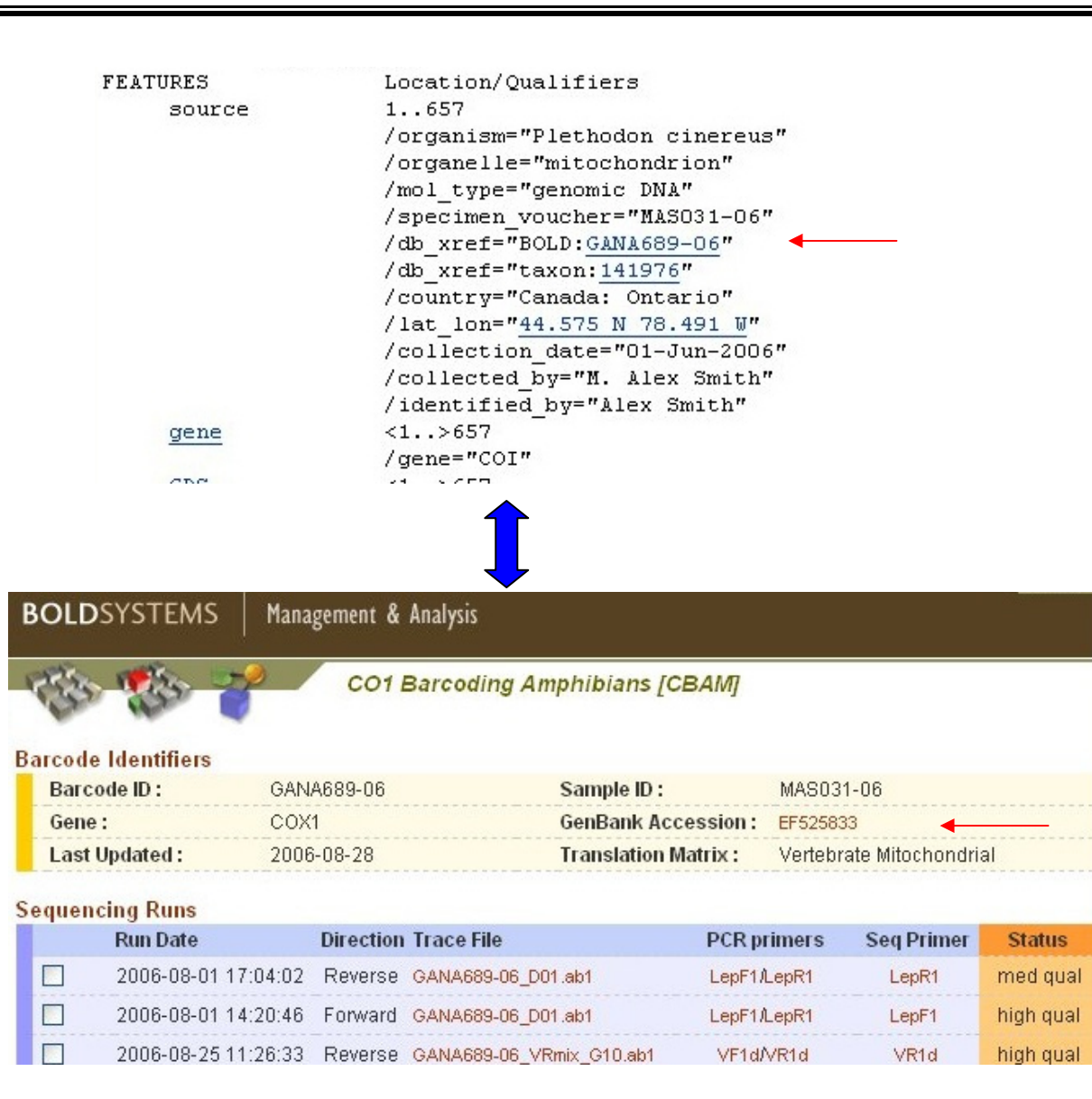
In addition to standardizing structure and content, database records need to be uniquely identified. This is demanded for internal management and for cross-referencing. For that reason, GenBank internally designates an 'accession number' to each set of sequence data plus corresponding annotations (Benson et al. 2008). This naming system is shared with the other INSDC members. The accession number originally comprised two letters from the Latin alphabet and five integers, but has been expanded to currently six integers, e.g. EU525415. It never changes over the whole lifetime of the record. If the sequence or the annotations are modified, the entry is given a new version number. In the GenBank flat file, information on the version can be

found in the line below the accession number (cf. Fig. 5). It consists of the accession number and an appended qualifier. In the exemplary record of Figure 5, this would be EU525415.1 for the first (unmodified) version of an entry. The gene index (GI) number also gives (indirect) versioning information: in the given example (Fig. 5), the gene index number GI:169882570 corresponds to EU525415.1. Accordingly, a new GI number would be assigned if a second version (EU525415.2) for the record was created. If a protein translation exists, it is assigned individual version numbers, as well (e.g. protein ID ACA97302.1 with the gene index GI:169882571). Thereby, the current version of an entry is always retrieved by its accession number, while its precursors are kept under their respective version numbers. BOLD features two types of records requiring unique identification: the sequence record and the specimen record (see Fig. 13). The sequence ID, which varies in length, consists of the acronym of the project the sequence was generated in and an appended numeric code (assigned internally), e.g. MHAHC825-05. The last two digits denote a date (in the example, the sequence information was compiled in the year 2005). Specimen IDs are not generated within BOLD, as they are the result from the process of voucher deposition. An acronym of the holding institution and its voucher number are adopted by BOLD (e.g.: 02-SRNP-23406 identifies a specific voucher held at the University of Pennsylvania). Barcode ID and specimen ID in BOLD are linked internally through hash tables. GenBank accession numbers are connected to the barcode ID in the same way. In GenBank, the BOLD barcode ID (as well as the specimen ID) are stored as source modifier directly within the respective object (cf. Fig. 11). Through such a form of cross-linking, if opening a GenBank record associated with a specific BOLD entry (e.g. GenBank accession number EF525833), one can navigate directly into the corresponding BOLD record and vice versa. Both GenBank and BOLD implement identifiers.

***Concerning the form of the identifier, GenBank seems to have found better traceable solutions,*** as it can accommodate information on previous record versions. Since only a single version can exist in BOLD, changes are of superimposing nature. Under the current system, new record versions in BOLD would in theory have to be listed under a new barcode ID (which would create confusion). BOLD should also link to the accession plus version number in GenBank and not the pure accession number (cf. Fig. 11). Especially in records that were migrated from GenBank and that have no original counterpart in BOLD.

***Concerning referencing through identifiers, GenBank records are more sustainable for molecular purposes, but are less informative and lack modularity.*** GenBank achieves a clearer link through directly integrating voucher number and specimen information into the sequence record. Nevertheless, BOLD can handle more specimen information by keeping it in a separate record. An additional benefit in this is that specimen information is not dependent on the barcode

Figure 11: Cross-referenced GenBank BARCODE and BOLD records



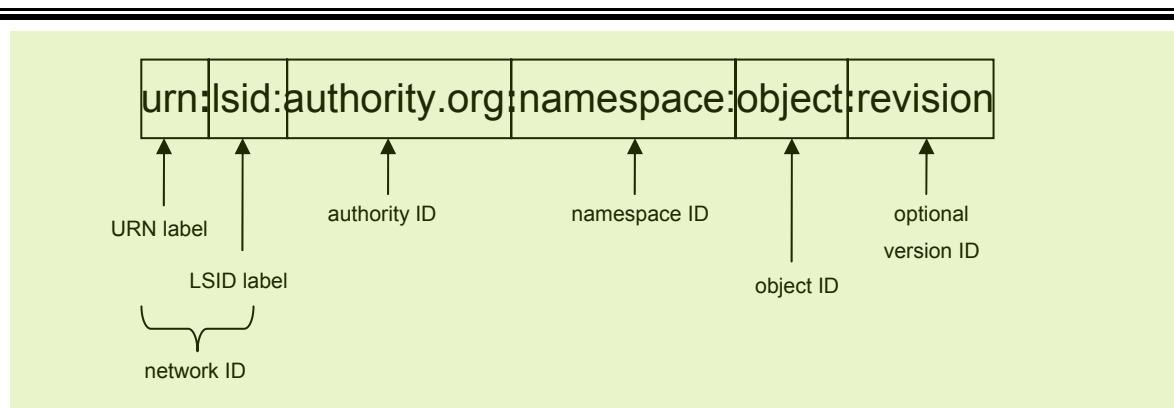
Compilation of a BOLD record and corresponding GenBank record. Red arrows indicate the respective link to the other database, which is realized through the Barcode or accession ID.

record and can be re-used. In the modularity of BOLD's approach, the specimen can become the centerpiece for integrating multiple kinds of information.

In the long run, DNA taxonomy data will be integrated with many other biodiversity data across different databases (see chapter 2.7). To achieve this, more identifiers will be necessary than just DNA sequence accession numbers or voucher numbers. Through identifiers, specimens will be linked to physical repositories, to DNA sequences, to digital photographs and other data

on the specimen (e.g. sound recordings, audiovisual sequences, etc.). They will formalize their taxonomic identification through connecting – via ID – to a central species registry (e.g. the above-mentioned ZooBank). This data cluster centering on species identity will be linked to several distributed specimen-independent data sources that give general information on the species (e.g. its ecology, literature on the species, etc.), enabling deeply structured queries for differentiated knowledge generation. This vision of the not-so-far future is likely to be realized by implementing Life Science Identifiers ([LSID](#); Clark et al. 2004; Page 2006). What are these? Like [DOI](#) and, to a lesser degree, like [PURL](#) in the publishing, library and similar fields, LSID are the form of unique identifier advocated for biological sciences by the TDWG (see chapter 2.3) and by GBIF (see chapter 2.7). Life Science Identifiers represent a special form of Uniform Resource Names (URN) and have been adopted by the Object Management Group ([OMG](#)). They are globally unique, persistent identifiers that serve the purpose to connect biological data through the Internet. Unlike the short-lived (Kahle 1997; Koehler 2002; Dellavalle et al. 2003), but directly applicable Uniform Resource Locators (URL), URN are independent from location. Instead, they depend on a namespace (and ID number; cf. Fig. 12) issued by a naming authority, for which neutrality in assigning names should be a necessary precondition. As URN, the success of LSID depends heavily on standardizing an LSID-adapted resolution process (apart from the guarantee of future scalability). By themselves, URN constitute no more than a name and cannot be entered into a regular browser (until 'armed' with an appropriate [extension](#), like those available for the Mozilla Firefox Internet Browser). The TDWG offers such a [resolver](#) that translates LSID into URL. In addition to the scenario depicted above, LSID can also be a way to merge distributed data that do not require an association with a formal taxonomic name (EDIT 2007). This is a promising path for some aspects of biodiversity (e.g. for the massive data input on environmental samples), but needs to be trodden carefully due to its abstractness. A problem concerning LSID as well as scientific names is the variability of species that the specialized taxonomist accepts as given and unavoidable, but that the end user of biodiversity data commonly forgets when considering 'species' as unmovable, stable entities (Berendsohn 1995; Krichevsky 2005). The elusiveness of species is maybe the most serious weakness in a comprehensive LSID system, since LSID have to identify unchanging targets. A way out of this dilemma would be to identify different taxonomic concepts (see TCS in chapter 2.3) through LSID and to deal with several coexisting concepts (EDIT 2007). However, species concepts and species criteria are themselves variable. A LSID network will have to be flexibly designed to meet taxonomic changes.

Figure 12: LSID syntax



The network ID at the beginning of the LSID is comprised of the urn:lsid: labels; authority ID is usually the root DNS of the authority issuing the LSID; the object ID is unique within the respective namespace ID; 5. an optional Revision Id to represent versioning information. Source: <http://lsids.sourceforge.net/>

*Currently, neither GenBank nor BOLD implement Life Science Identifiers.* The discussed DNA sequence databases will be a small but important component in such a global identification and expert system (in one) and will have to achieve compatibility through adopting LSID. Especially so if BOLD rigs up and tries to integrate more of the involved components into its system – a possibility that today does not seem impossible due to the mentioned modularity of BOLD and its structure: the Barcode of Life Data System is inclusively a nucleotide database. GenBank is so exclusively. It is the database component of another complex data system – that maintained by NCBI –, but which focuses on molecular biology in general and not on biodiversity. At the latest when adopting LSID, which identify unchanging objects but can indicate the version of an object, BOLD should start versioning its data.

## 2.5 Quality of the information

The previous chapter focused on interconnecting DNA taxonomy data. When increasing the availability of data and its re-use probability by creating such a network, it is relevant to have an idea of the quality of the shared data. When referring to data quality, I concentrate on data accuracy, which is the major issue in molecular biology database quality. It gains the more importance as the data 'flooding' continues and as errors are propagated and multiplied throughout the database by making deductions from erroneous records (Krichevsky 2005). Accuracy of submitted sequences can be very heterogeneous in databases – it reportedly is in GenBank, which acts as repository for all publicly available sequences (Brusic et al. 2000; Ruedas et al. 2000;



Bridge et al. 2003; Forster 2003; Harris 2003; Ross & Murugan 2006). Errors that can be transmitted to the database arise at various levels:

- while handling the field data, the laboratory data, or the sequence data (field or vial number errors, wrong metadata attached to a sequence, etc. – the last phenomenon can occur both before and after submission)
- through contamination of the sample, technical failure or signal erosion in the laboratory (more common in earlier times, sequencing errors became less significant in recent years; however, contamination and 'bad' sequences with many ambiguous positions are still a problem)
- associated with the identification of a specimen or sequence (the identification process or the identification source used in the process may lead to errors, as can a subsequent change in taxonomy).

First and foremost, the producing researcher is responsible for the quality of the data, as he/she often represents the only serious instance to control the data. Wouters & Beaulieu (2002) state that if "data sharing becomes increasingly mediated by information and communication technologies, and hence less dependent on face to face communication, the generation of trust [and of quality control] will have to be organised differently". The notion to introduce quality control through the journals in which the respective articles are published (Harris 2003) does not seem very sustainable, as not all submitted sequences are conventionally published (apart from the lacking resources in small journals and possibly the lacking willingness in big journals). Hence, secondary quality control has to be performed by the repository. GenBank already checks automatically for contamination of the sequence with another organism, correct orientation of the sequence and examines taxonomic assignment as well as bibliographic information. But from the above studies it can be assumed that this is not enough to exclude errors. Through its current focus on a single gene, BOLD has more options to check sequence consistency. Besides screening for contaminants (for typical sources like the analyst herself/himself), BOLD also verifies their origin from the COI gene through a Hidden Markov Model (Eddy 1998) and subsequently checks for pseudogenes<sup>7</sup>. Furthermore, it highlights sequences with more than 1% ambiguous bases for review. For each of its sequences, BOLD indicates the number of ambiguities they contain (cf. Fig. 13). Neither BOLD nor GenBank check for possible sequencing errors. Although redundancy is often criticized as a drawback in primary databases like GenBank (which is valid in other contexts and led NCBI to establish the derived, curated database [RefSeq](#)), in DNA taxonomy it is in fact needed. Many similar specimens have to be investigated to realize by which criteria they

---

<sup>7</sup> Pseudogenes are ancient copies of the target gene that were inserted elsewhere into the genome and that have lost their function. They may compromise the analysis when mistaken for the target sequence.

can be grouped together and what makes them different from other groups, i.e. where the species boundaries lie. This requirement in taxonomy partly helps to guard against database errors, as outliers will draw attention. The other requirement of taxonomy to always 'keep in touch' with the specimen and its collateral data can also be very helpful in cases that need to be re-examined. This need not necessarily be physically – often, the location data, collecting time or other details, especially photographs deposited in BOLD can give ample evidence (cf. Fig. 14). Examining the vouchers (Huber 1998, Agerer et al. 2000) is the only way to straighten out errors (or disperse doubts), but this is often impossible in other biological disciplines due to less rigorous vouchering. In addition to this, the standards to obtain barcode status (e.g. double stranded vs. single stranded sequencing) help in guarding against errors and CBOL is granted the right by GenBank to withdraw the BARCODE status of any record they consider unsuited (this, by the way, is a unique case in which somebody else than the submitter has influence on a GenBank annotation).

Figure 13: BOLD project record list indicating sequence ambiguities

<input type="checkbox"/>	Plethodon cinereus	MAS104-05	AMPAS164-05	657 [0n]
<input type="checkbox"/>	Plethodon cinereus	MAS314-05	AMPAS208-06	657 [0n]
<input type="checkbox"/>	Plethodon cinereus	MAS321-05	AMPAS209-06	657 [0n]
<input type="checkbox"/>	Plethodon cinereus	MAS031-06	GANA689-06	657 [0n]
<input type="checkbox"/>	Plethodon cinereus	MAS049-06	GANA707-06	<b>634 [7n]</b>
<input type="checkbox"/>	Plethodon cinereus	MAS133-05	AMPAS165-05	620 [0n]
<input type="checkbox"/>	Plethodon cinereus	MAS334-05	AMPAS210-06	611 [1n]
<input type="checkbox"/>	Plethodon cinereus	MAS070-05	AMPAS161-05	608 [1n]
<input type="checkbox"/>	Plethodon cinereus	MAS342-05	AMPAS211-06	601 [0n]
<input type="checkbox"/>	Pleurodeles waltl waltl	RuHF-PWWV-01-2	ABCAP102-05	657 [0n]
<input type="checkbox"/>	Pleurodeles waltl waltl	RuHF-PWWV-01-3	ABCAP103-05	657 [0n]
<input type="checkbox"/>	Pleurodeles waltl waltl	RuHF-PWWV-01-1	ABCAP101-05	656 [1n]
<input type="checkbox"/>	Pseudacris crucifer	MAS074-05	AMPAS105-05	416 [3n]
<input type="checkbox"/>	Pseudacris crucifer	HLC-10597	AMPAS187-05	657 [0n]
<input type="checkbox"/>	Pseudacris crucifer	MAS058-05	AMPAS104-05	657 [0n]
<input type="checkbox"/>	Pseudacris triseriata	HBL008477	AMPAS184-05	419 [0n]
<input type="checkbox"/>	Pseudacris triseriata	MAS152-05	AMPAS189-05	419 [0n]

BOLD specimen list for a specific project. Sequence length is given in the last column. Square brackets behind the sequence length indicate the number of ambiguous positions in the respective sequence. Red, bold: more than 1% of positions in this sequence are ambiguous (i.e. do not show a clear signal).



Figure 14: BOLD specimen page with photograph


Birds of North America [TZBNA]

---

### Specimen Identifiers

<b>Sample ID :</b>	CWW 1260	<b>Museum ID :</b>	CWW 1260
<b>Isolate / Field Num:</b>		<b>Collection Code :</b>	
<b>Donated By :</b>		<b>Deposited In :</b>	Royal Ontario Museum

### Taxonomy


<b>Identifier :</b>	
<b>phylum :</b>	Chordata
<b>class :</b>	Aves
<b>order :</b>	Anseriformes
<b>family :</b>	Anatidae
<b>genus :</b>	Chen
<b>species :</b>	<i>Chen caerulescens</i>

### Specimen Details

<b>Voucher Type :</b>	
<b>Tissue Type :</b>	
<b>Extra Info :</b>	Chen caerulescens
<b>Sex :</b>	U
<b>Reproduction :</b>	
<b>Life Stage :</b>	
<b>Note :</b>	


### Collection Data

<b>Collectors :</b>	
<b>Date Collected :</b>	
<b>Country :</b>	Canada
<b>State/Province :</b>	Nunavut
<b>Region/County :</b>	Hannah Bay
<b>Sector :</b>	
<b>Exact Site :</b>	
<b>Latitude :</b>	51.211
<b>Longitude :</b>	-79.81
<b>Coord. Source :</b>	
<b>Elevation/Depth :</b>	



### Photographs

Lateral



BOLD specimen page connected to a barcode record. In addition to collecting locality, the picture of the specimen can help resolve 'suspicious' cases without necessarily accessing the cryoconserved voucher (e.g. tissue/blood or DNA).

Thus, *BOLD and to a lesser degree also BARCODE entries in GenBank show an increased 'impermeability' against database errors and stand a better chance to correct them.*

Besides, arguing philosophically, the application of molecular methods in order to solve taxonomic questions seems to (but need not necessarily) render a better protection from misidentifications than when setting out to answer completely different questions. Nevertheless, neither GanBank nor BOLD are manually curated databases other than through the submitters themselves (unlike e.g. the [SwissProt](#) protein sequence database) and serious errors in sequence or annotation can occur. Although highly desirable, curation of BARCODE records would probably lead GenBank onto a slippery slope, as it serves a multifaceted community. BOLD, however, should consider this option as soon as enough funding is available.

## **2.6 Preservation of molecular taxonomic data**

The previous chapter focused on quality in terms of accuracy, but since long-term availability of DNA taxonomy data is also a 'requirement', it can be seen as another aspect of quality (cf. ISO 9000). Now how do we arrange for the continuance of these data? As pointed out already, DNA taxonomy information exists overwhelmingly as born-digital data, which means that they are originally digital and – extrapolating at the moment of production – not meant to exist in analogous form (the 'tip of the iceberg' accession numbers are often published in scientific journals, but this does not imply transcending the digital medium). Although it is usually not publicly realized, the preservation of digital data is a challenging and pressing issue (cf. Borghoff et al. 2003, 2005; Schwens & Liegmann 2004; nestor 2006). The UNESCO perceives an acute danger of losing collective memory and cultural identity (UNESCO 2003) through unpreserved digital information. And referring to the volatility of information on the Internet (cf. Kahle 1997; Koehler 2002; Dellavalle et al. 2003), IT pioneer D. Hillis coined the term of the "digital dark age" for the 21<sup>st</sup> century.

Generally, the problem of digital preservation has several layers:

- the necessary dimensions of preservation in terms of data volume
- the heterogeneity of the objects to be preserved
- the short lifetime of most digital objects
- the influence of legal concepts

How do these aspects apply to DNA taxonomy? Concerning the *amount* of data: a study at the UC Berkeley (Lyman & Varian 2003) and following market research studies revealed that most of our 'information' exists in digital form and that currently, several exabyte of data are produced

yearly (1 exabyte =  $10^{18}$  byte). The problem associated with such astronomically high data volumes is how to select the relevant information that can be preserved realistically. I argue that at this moment, no specific records should be selected for preservation among the annotated DNA sequence data in GenBank and BOLD. (In any case, selection criteria would be very difficult if not impossible to establish due to the pluralism in molecular biology interests). Unlike other data in molecular biology (structural, simulations, etc.), the textual data describing DNA sequences take up very little storage space. For instance, the sequence information included in the last GenBank release (NCBI-GenBank Flat File Release 165.0, April 2008; without whole genome shotgun sequences) occupies only somewhat more than 300 Gigabyte. This volume would fit on the hard drive of a contemporary personal computer<sup>8</sup>. With the current developments in computer industry, it does not seem easy to accumulate as much sequence information as to head into a storage problem. A different situation arises when considering trace files (cf. Fig. 4): a single typical sequence trace file for the COI gene occupies slightly more than 200 Kilobyte, about half as much if compressed. If we assume two compressed trace files per sequence (it probably is overly optimistic for taxonomy to assume that much), this would make around 80 gigabyte for the 400,000 barcode records presently contained in BOLD. This, also, is still a manageable volume. Although barcode-'harvesting' through [iBOL](#) and other big science projects is going to scale up considerably (cf. Hajibabaei et al. 2005, EDIT 2007), even current digital storage technology is able to cope with the anticipated data volume (BOLD has funding to increase its current 5 terabyte of regular storage capacity to more than 70 terabyte; Ratnasingham & Hebert 2007). The same applies to non-molecular data included in BOLD, e.g. digital images (Ratnasingham & Hebert 2007 calculate with capacity to store and analyze 10 million records in BOLD).

Hence, *barcode data volume is no problem if extrapolating from current methods*. However, new methods are likely to be introduced over time and it cannot be precluded that data storage alone might turn into a problem at some point. This seems unlikely for genome data, as the same as above applies: they consist mostly of textual data (e.g. the human genome takes up less than a gigabyte and it is one of the larger of the several hundred genomes sequenced so far).

Concerning the *heterogeneity* of the objects: digital archives ('archives' in the sense of library and information science, not computer technology) are continuously confronted with the necessary problem of having to manage a plethora of different document types (again, in the LIS sense), each with its own internal structure. One possible approach to meet this difficulty is to select and preserve only specific file formats. Thus, only a single or a few 'survival plans' have to be worked out and optimized for the reduced range of documents (this approach was chosen e.g.

---

<sup>8</sup> But the string of bases it codes for would be enough to be wrapped around the Earth more than 30 times (if we assume that 10 base pairs take up around 2.3 cm).

by the [Florida Digital Archive](#)). For the highly specialized information in DNA taxonomy, again, no selection becomes necessary, as presently, the same technique is consistently applied in parallel manner (DNA sequencing). Thanks to this, the properties and preservation needs of the resulting data are identical, which bodes well for their preservation. However, many different ways to logically store the same data are possible and only one format should be chosen as standard. As the databases under study not only store their records but have them accessed, searched and analyzed regularly by users, each stores its data in an internally standardized way.

*GenBank and BOLD manage only a single format for their genetic data.* BOLD also contains trace files (as does the NCBI [Trace Archive](#)) and photographs, each in one format only (but maybe not particularly well chosen for preservation: the commercial ab1 for traces and the lossy jpg for pictures).

Questions concerning the *lifetime* of objects in DNA taxonomy depend on the view one adopts of the digital object (Thibodeau 2002): as physical object, i.e. as bit stream stored in a defined location, we are interested in its integrity (and authenticity), but not so much in its absolute longevity after the object has been replicated or refreshed. At the logical layer, digital objects interpret the physical object through specific data formats. It is desirable, but not necessary that these formats have a long lifespan. However, after the lifespan of a logical object expires due to contingencies of the IT environment, its content can be migrated by transforming its logical structure, i.e. by converting it into a new format. By doing so effectively, we still retain the conceptual object that can be understood by a person or knowledge-based system. It is the lifetime of the conceptual object and its "significant properties" (Thibodeau 2002) that we should be mainly interested in. The easiest way to preserve the conceptual object would be to preserve the underlying physical and especially the logical layer. Simple media migrations or refreshments against physical deterioration can effortlessly be carried out on DNA taxonomy data owing to their manageable volume and no special considerations are necessary here, as long as the integrity of the bit stream is not compromised during this process (algorithms can check this and GenBank probably makes use of these; when the need for media migration arises for BOLD in the future, it should do so as well). Content migrations are more complicated and error-prone. Thus, formats should be chosen that are likely to be supported as long as possible (e.g. through wide application and flexibility of the format). This applies to GenBank's data description format ASN.1, and probably also to BOLD, which I suppose uses some form of XML (no exact information could be obtained). Both formats are standardized and independent (i.e. not controlled by a company), modular (separating content and structure) and are machine- and human-readable (helpful in case the conversion fails...). They also have the advantage that their structure specification and the associated metadata are stored internally (I presuppose this for BOLD). Both are more or less well

documented, although information on ASN.1 is scarcer (it is symptomatic that ASN.1 is not listed in the [PRONOM](#) file format registry). The last aspect (along with the modularity aspect) is of relevance when applying an emulation strategy. For this purpose, the databases should keep copies of their previous formats when forced to migrate in the future (GenBank should also convert to XML, which it already does if a user demands an XML output). However, such an approach would also have to address the highly complex task to emulate the corresponding database, so that emulation is not likely to be a promising preservation strategy for DNA taxonomy. Without maintaining the functionalities around the records, the preservation of these as conceptual objects would be impaired: the significant properties of our objects imply offering sequence information data and corresponding annotations in comparative form, i.e. linked to its equals<sup>9</sup>. Nevertheless, the individual records should also be stored separately outside the database system so that in the worst case the functionalities could be rebuilt around them. GenBank already implements this, but BOLD should also store its data in database-independent form, e.g. on its backup tapes. One could also consider storing such data in immediately human-readable, more permanent physical form like microfilm. With the current data volume, this would still be feasible and would create a security beyond that offered by regular digital preservation ("regular", as microfilm is and especially will be machine-readable as well).

So far, I have only considered intrinsic factors that influence the objects' lifetime. But extrinsic factors also apply, especially the question of database funding. Among others, this topic is addressed in the inter-union bioinformatics group report (Berendsen 2003). This group (1998-2002) had the aim to bring the availability of molecular data into public attention. Unfortunately, it had little noticeable impact (B. Robson, advisory board, IBM, pers. comm.), maybe due to its limited dispersal (A. Lesk, representing CODATA, pers. comm.). The report draws attention to the fact that all funding models of the primary databases operate on a short-term basis, with no international commitment (from countries other than the US, the EU, and Japan) and that such a system, based on the goodwill of the individuals currently involved, is a threat to the integrity of those databases. While indeed the situation should change for the primary databases, BOLD cannot (yet?) expect such a form of support. Instead, it currently depends on Canadian (and partly US American) funding which will end by 2011. What will happen afterwards? Apparently, the University of Guelph (where BOLD is hosted) has recently agreed to provide support to maintain the existing infrastructure and a minimal team. This would not allow any developing work but would guarantee the persistence of the system. Beyond that, the International Barcode of Life

---

<sup>9</sup> The boundaries of our conceptual objects blur, as we are dealing with individual entities, but with entities that are all merged in a functional network (much like the 'individuals' in a living coral). A network, by the way, that relies on the Internet.



Initiative ([iBOL](#)) will support BOLD with almost the 15th part of the funds raised during the next years (S. Ratnasingham, pers. comm.).

*Owing to the fact that DNA taxonomy data reside in databases and that the functional framework offered by these defines our notion of the data, conservation is not trivial technically. On the level of the supporting organizations, more stability is necessary.* Regarding the latter, it would be highly desirable if especially GenBank, but also BOLD would be allowed to and also committed themselves to operate in a context demanded for 'trusted digital repositories' (RLG 2002). Related to this, both databases should consider adopting – as far as the transactional database structure allows – the ISO-standardized Reference Model for an Open Archival Information System ([OAIS](#)). BOLD already "implements certain characteristics of the OAIS" (S. Ratnasingham, pers. comm.), but GenBank does not commit itself to it (nevertheless being called an 'archival' database). BOLD and GenBank (individually, not only as NCBI component) should develop explicit mission statements, specifying, among others, the needs of their target audience and how they plan to meet changes in these needs. Although for the wider community the proactive preservation of DNA sequence data is especially important at NCBI (centralization of data, cf. Benson et al. 2008), it would be beneficial for taxonomy if BOLD, in cooperation with GenBank, also fulfilled an archival function. BOLD holds more (and deeper structured) data and a certain degree of redundancy would bestow increased safety upon preservation efforts.

Concerning *legal aspects*: often, digital preservation efforts are constrained by digital rights management (DRM) obligations. Some users may have restricted access to specific content, or licenses may have to be considered, possibly impeding even the simplest form of migration. Luckily, copyright is not directly an issue in public nucleotide sequence databases (although GenBank includes sequences from issued patents, cf. Fig. 15). Therefore, no administrative metadata on DRM are necessary. All sequences in GenBank and BOLD are freely accessible (excepting those in the BOLD LIMS that are being prepared for publication). This should not be considered as granted: in 1996, the World Intellectual Property Organization ([WIPO](#); Geneva, Switzerland) published a draft that would have extended copyright agreements to databases (cf. Berendsen 2003). The draft was not adopted, but the incidence illustrates the vulnerability of such data (cf. Garnier & Berendsen 2002).

In DNA taxonomy, possibly some information regarding collecting locality may have to be withheld in some instances where the population or species is endangered. But rather than introducing DRM to suit this scenario, the rare cases where this would be necessary should be handled in another way. For example by not resolving locality information below country level and obtaining permission from CBOL to nevertheless submit the record formally as 'barcode'.

GenBank and BOLD consider the sequence records as property of the submitter(s), who are the only users allowed to make modifications to the records. This increases stability, but impedes 'social annotation'. NCBI offers a database of Third Party Annotations ([TPA](#)), but these are only released by NCBI after publication in a peer-reviewed journal.

*As BOLD and GenBank offer free content, digital preservation is currently not constrained by considerations on rights management.*

Figure 15: GenBank sequence relating to an issued patent

---

```

LOCUS       CS742451                201 bp    DNA     linear   PAT 08-MAY-2008
DEFINITION  Sequence 10447 from Patent WO2005083127.
ACCESSION   CS742451
VERSION     CS742451.1  GI:187841478
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Luke, M.C., Devlin, J.C. and Cargill, M.C.
  TITLE     Genetic polymorphisms associated with stroke, methods of detection
            and uses thereof
  JOURNAL   Patent: WO 2005083127-A 10447 09-SEP-2005;
            Applera Corporation (US)
FEATURES             Location/Qualifiers
  source             1..201
                    /organism="Homo sapiens"
                    /mol_type="unassigned DNA"
                    /db_xref="taxon:9606"
ORIGIN
  1  gaggatgggg  cggttggcct  ggcattgagtg  ttgaaccaga  aaatgggcct  ggggagggca
  61  gagctggaga  cactttgaac  gccatgcttg  gtaggtgtgg  raatggggac  gcgttctgtt
 121  cagaggtcat  cccggaagcc  tgccgtgtgc  agactggagg  cagggaggat  tgtttgaagg
 181  ttacgcaaga  gtccaggcac  a
//

```

---

Specification of the patent and the patent holder is given in the 'Journal' field. Source:

<http://www.ncbi.nlm.nih.gov/Genbank/>

---

## 2.7 Data integration

With the breakthrough of information and communication technologies in systematics (Hine 2008), a "quiet revolution" (Bisby 2000) remodels the field of biodiversity and turns most of its questions into questions of biodiversity informatics. In chapter 2.4, I already roughly

outlined a model of how biodiversity information might be interconnected: a data 'cluster' oriented at the specimen (specimen voucher, sequence, picture, audio, video, etc.) would be linked, via LSID specifying the taxonomic placement, to a heterogeneous pool of data on the specific taxon (cf. Edwards et al. 2000). This data pool would provide information related to, for example, species description, ecological parameters, ecosystems, phylogeny, evolution, conservation, geospatial aspects, climate, taxon specialists, current studies/projects on the taxon, etc. The seamless integration of these sources is only possible through the Internet, offering the human user a single, centralized point of access and permitting the partly automated generation of synthetic knowledge (or even fully automated through 'knowbots', Frishman et al. 1998). Automated output could focus, among many other possibilities, on identification keys for specific taxa, on natural resource management incentives, on priority lists for conservation biology (in terms of taxon or geographic area), on analyses of climate influence, on knowledge gaps and on incentives for research funding. Part of this scenario is already being put into practice. There exist a confusing multitude of individual projects with limited scope (e.g. databases focusing on a specific group of organisms, like [ants](#), or [algae](#)). These constitute the individual elements of a comprehensive biodiversity resource, but the real challenge lies in integrating them (Bisby 2000). The [Species 2000](#) database federation (brought into life by the Committee on Data for Science and Technology, [CODATA](#), and the International Union of Biological Sciences, [IUBS](#)) compiles the 'Catalogue of Life' ([CoL](#)) and is aided in this task by the Integrated Taxonomic Information System ([ITIS](#)). At present, the Catalogue of Life is the most complete index of scientific species names (as well as of synonyms and common names). Of the 1.7 to 1.8 million described species, already more than 1.1 million are covered in this catalog. Currently, more than 50 individual taxonomic databases contribute to this number. The Catalogue of Life is used, among others, by the recently available Encyclopedia of Life ([EOL](#)), an enterprise that tries to fill those names from the catalog with content on the respective species (cf. Fig. 16). The Encyclopedia of Life is a realization of systematist E.O. Wilson's vision to procure information on each individual known species (Wilson 2003). Currently, only about 30,000 species pages of the encyclopedia are associated with additional information (Kelly 2008), but the partly aggregating, partly (mediated) collaborative 'Web 2.0' approach chosen by the organizers (Harvard University and partners, among others the Biodiversity Heritage Library, [BHL](#)) warrants considerable increase in species data (especially if building up a connection to [Wikispecies](#) and using synergies). The Encyclopedia of Life focuses on a wide spectrum of species-describing data and on species-related literature. Besides also considering species data, the international, inter-institutional Global Biodiversity Information Facility ([GBIF](#)) centers primarily on individual biodiversity observations and specimen data, accessing the whole wealth of natural history collections data



Figure 16: Exemplary record obtained from the Encyclopedia of Life

HOME
PREFERENCES
LANGUAGE: EN
FEEDBACK
PRESS ROOM
USING THE SITE
ABOUT EOL

You are not logged in. Please [login](#) or [create an account](#).

## IMPERIAL BLUE BUTTERFLY

*Jalmenus evagoras*

IUCN RED LIST STATUS: NOT EVALUATED

CLASSIFICATION : [TEXT](#) | [GRAPHIC](#) | [SOURCE](#)

IMAGES
MAPS

IMAGES

PAGE 1    NEXT >

Image is Some rights reserved  
 AUTHOR: [Martin Purvis](#)

Adult male upperside

**Animals +**  
 Arthropods +  
 Insects +  
 Butterflies and moths +  
 Butterflies +  
 Blues +  
 Jalmenus +  
**Imperial Blue Butterfly**

Archaea +  
 Bacteria +  
 Chromista +  
 Fungi +  
 Plants +  
 Protozoa +  
 Viruses +

LESS    **DETAIL**    MORE

TABLE OF CONTENTS

- Overview
- ▶ Introduction
- Description**
- ▶ Physical Description
- ▶ Succinct
- ▶ Etymology
- ▶ Identification
- ▶ Chromosomal Data
- ▶ Original Description
- Ecology and Distribution**
- ▶ Distribution
- ▶ Habitat
- ▶ Host Plants
- ▶ Attendant Ants
- ▶ Phenology
- ▶ Mating
- Conservation**
- ▶ Trends and Threats
- Evolution and Systematics**
- ▶ Classification
- ▶ Paleontology
- ▶ Nomenclatural History
- ▶ Concepts and Synonymy
- Relevance**
- ▶ Culture
- References and More Information**
- ▶ Literature References
- ▶ Editor's Links
- ▶ Specialist Projects
- ▶ Common Names

**INTRODUCTION**

---

SOURCE AND ADDITIONAL INFORMATION  
[Rod Eastwood](#)  
 Some rights reserved  
 (cc) BY

*Jalmenus evagoras* is a medium sized lycaenid butterfly endemic to south eastern Australia. Adults have iridescent greenish blue wings bordered in black with a long thin tail. Underside is yellow-buff with a series of black spots and bands. Larvae are dark coloured and have a row of short fleshy projections each side of the dorsal ridge. They are gregarious and feed openly during the day on a variety of *Acacia* plants where they are attended by swarms of small black *Iridomyrmex* ants. Adults of *J. evagoras* can occur in large populations but are usually localised throughout the species range. This patchy distribution is in part due to the dependence of *J. evagoras* upon the overlapping requirements of suitable host plant and attendant ants. Ants are attracted to sweet secretions produced by the butterfly larvae and in return they protect the larvae from a range of predators and parasitoids. *J. evagoras* has been used as a model organism for testing a range of behavioural and evolutionary hypotheses. See [Pierce lab website](#) for list of publications.

EXPLORE

[Raja miraletus Linnaeus, 1758](#)  
 Brown ray

[Barbus oxyrinchus Pfeffer, 1839](#)  
 Pangani barb

[Heteroleotris tentaculata \(Smith, 1958\)](#)  
 Locusthead

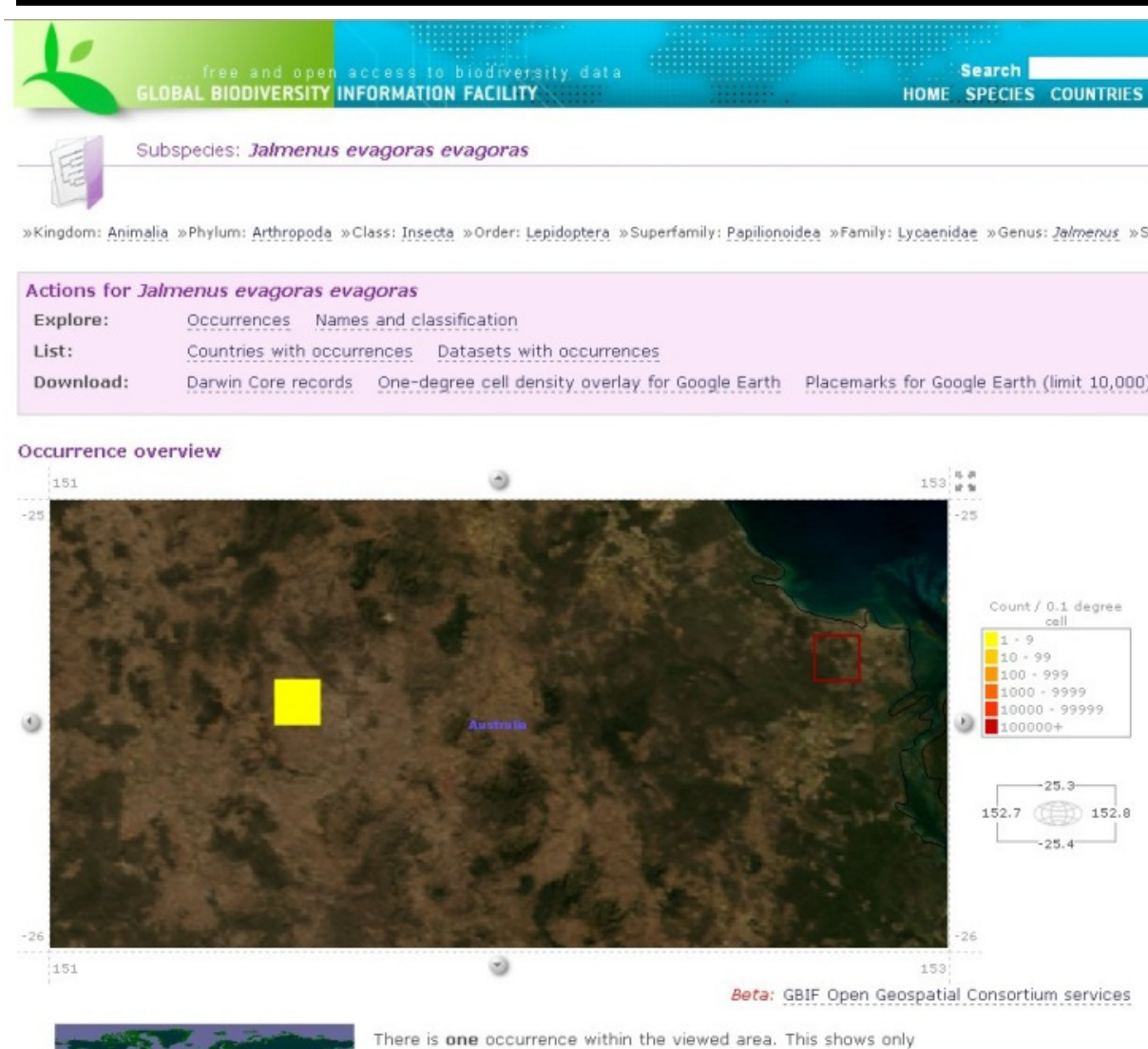
[Saurida flamma Waples, 1962](#)  
 Orangemouth lizardfish

[Enthrocles schlegelii \(Richardson, 1846\)](#)  
 Japanese rubyfish

Example of one of the pages in Encyclopedia that are already connected to additional species information.

Source: <http://www.eol.org/>

Figure 17: Exemplary GBIF record



Typical GBIF output. The GBIF query was conducted on the same species as that for the Encyclopedia of Life in Fig. 16 (in EOL, it was picked from the 'model' species pages). Note the 'Occurrences' link (in this case, it would link to museum specimens) and the Darwin Core download option. Source: <http://www.gbif.org/>

(Brooke 2000). As pure data aggregator, GBIF lacks a creative approach like that of the Encyclopedia of Life, but is instead more concerned in how to interconnect hundreds of millions of biodiversity records (observations, statistics, collection data, etc.). Consequently, more weight is put into machine-readability and the application of data exchange standards (like e.g. Darwin Core; cf. Fig. 17).

Where do molecular data fit in? The Encyclopedia of Life offers bibliographic data for molecular studies as well as manually selected GenBank sequence accessions, which are

imbedded in strings of natural language and without direct links (cf. Fig. 18). GBIF states that it will refer to molecular information in the future (Canhos et al. 2004). So not much has as yet happened in integrating molecular information into the global biodiversity information framework. Care has to be taken not to duplicate efforts – a serious danger constantly looming in the complex and unstructured biodiversity landscape (experience has shown that this happens both at the underlying and at the integrating levels). How is this to be prevented? Eclectic approaches to manually compile sequence registers that lack any computer-interpretable structure (like the one in Fig. 18) are no sustainable solution<sup>10</sup>. A direct cooperation with GenBank or

Figure 18: Sequence information as shown in an Encyclopedia of Life record

Description	
▶ Physical Description	<i>Jalmenus evagoras</i> Mitochondrial Cytochrome Oxidase subunit 1 (COI) DNA sequences available from <a href="#">GenBank</a> : DQ249942 to DQ249949, DQ249952. Sequence data (COI) also available for attendant <i>Iridomyrmex</i> ants: DQ249954 to DQ249980, DQ249982, DQ249983, DQ249985 to DQ249989.
▶ Succinct	
▶ Molecular Biology and Genetics	

Section of the same Encyclopedia of Life record as shown in Fig. 16. Source:

<http://www.eol.org/>

BOLD is necessary. First, these databases will have to clarify *who* is going to deliver *which information* to *whom*. '*Who*'? It seems sensible to always procure the GenBank accession numbers in any comprehensive biodiversity information resource. However, this can be supplied by BOLD. In fact, I would argue that the orientation and structure of BOLD make this entity more suited to interact with the biodiversity community. BOLD focuses more on the specimen than GenBank does, and it keeps the relevant data ready in the corresponding structure. Furthermore, its data sharing and processing tools that transcend the usual database functionalities might become important on a more global scale in the future. '*Which information*'? COI barcodes seem the most sensible data to start with, as they are currently collected in a pattern that meets the all-species notion of a global biodiversity platform. Further genes can follow, and [alternative markers](#) should be offered in groups where COI fails to resolve species boundaries. Along with a stable

<sup>10</sup> Why should not Encyclopedia of Life search for sequences on demand? This is implemented in another realization to have a page on each species: R. Page's [iSpecies](#) script creates an ad-hoc record for a queried species name and delivers results, among others, extracted from Yahoo image and Google Scholar searches. Its only deficit is that it is dependent on how the taxon name is entered: i.e. synonyms are not resolved (e.g. through access to the Catalogue of Life) and misspellings are not coped with (e.g. through fuzzy searching). But see Page (2005).

link to the sequence information (e.g. accession number integrated into LSID), metadata should be communicated to the derived database in order to offer further information retrieval criteria. '*To whom*'? At this moment, GBIF and the Encyclopedia of Life seem to be the only sensible recipients (as some other projects, like e.g. the [All Species](#) Foundation seem to lack either a profile that contrasts clearly with those initiatives named above and/or the necessary perspective for continuation). The Encyclopedia of Life has as yet not introduced a rigorous data structure, so that its function in a superimposed knowledge base seems less significant than that of GBIF, which is also wider in scope. Hence, BOLD should interact primarily with GBIF, but should offer the GBIF-tailored records also to the Encyclopedia of Life. To meet the increased access that would result from integration into a biodiversity portal, BOLD will have to develop mirror sites. The aim of BOLD is to be used as a general species identification system via barcodes, so that usage is expected to increase drastically. BOLD is therefore already investigating different mirror models (S. Ratnasingham, pers. comm.). Developing regional nodes that derive from the BOLD structure (like the one at the Bioinformatics Center in South Korea that is to be launched soon) will also aid in decentralizing the access<sup>11</sup>.

These are very theoretic considerations, and time might introduce further variables into the 'equation'. As a matter of fact, although we already possess the technological key to a global, integrated biodiversity information system, the dynamics involved in establishing it still leave many questions open and we are still in the dark as to which specific form such a system will adopt. It is likely that a future comprehensive biodiversity portal will do more than just retrieve information (or generate knowledge based on this information). Probably, it will involve also a workbench for submitting a variety of data types – out of convenience, but in times of dwindling taxonomist numbers also a necessity (in order to integrate citizen scientists). BOLD should not wait too long with adapting its LIMS to such an emerging system. Eventually, specimen information for ongoing projects might be entered, managed and shared centrally and BOLD might want to contribute in shaping the necessary interface, adapting it to the needs of barcoding as part of an integrated taxonomy. However, what so far has been mostly left out are formalized requirements analyses for biodiversity portal initiatives (Neale et al. 2007). Much more weight should be given to these – globally but also at the level of individual databases (neither GenBank nor BOLD have performed such analyses). Biodiversity portals have to be constructed around user needs. And the user should not be seen as distinct from data providers, but as one of them.

---

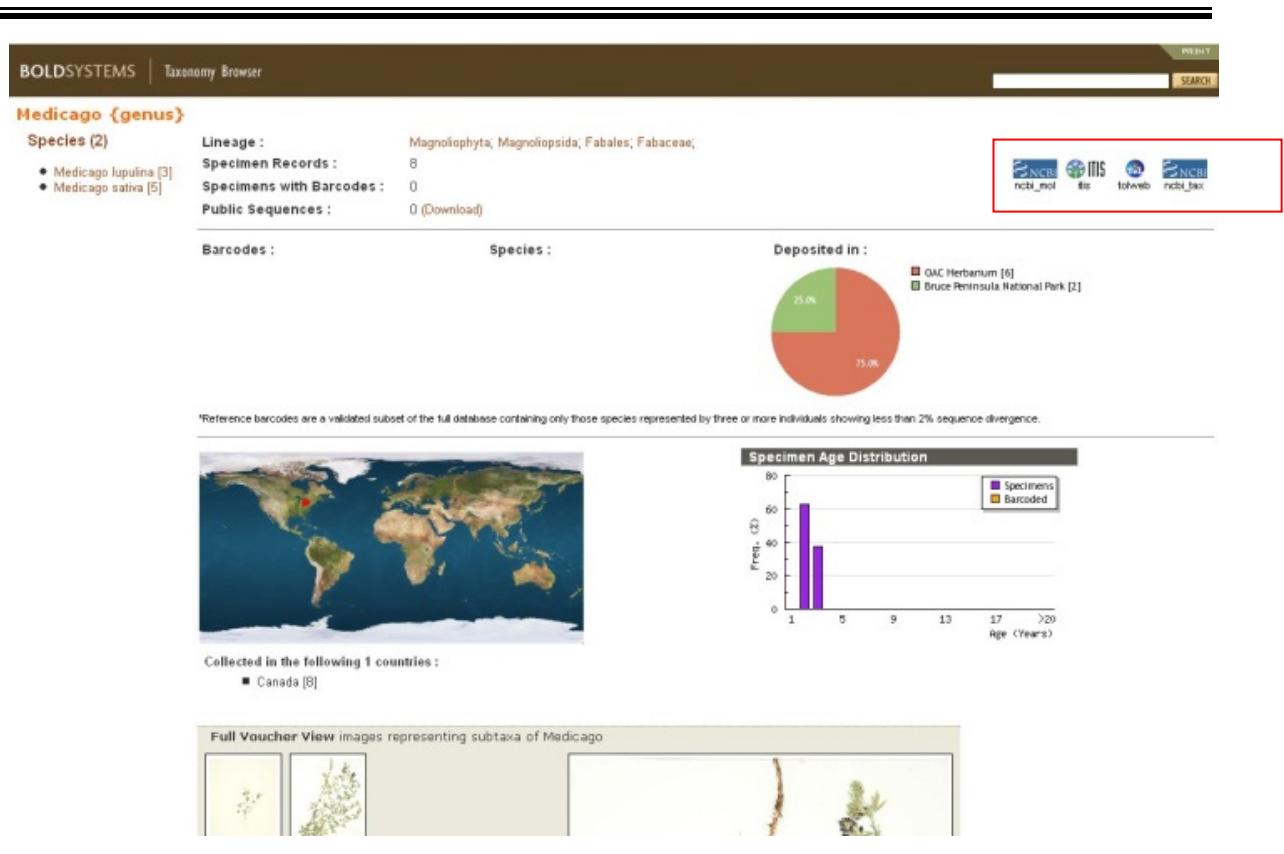
<sup>11</sup> Real decentralization will be achieved if the barcoding 'client' of the future, e.g. a sequencing device carried into the field (e.g. Blazej et al. 2006) that is connected to the Internet pre-sorts the evidence based on a hypothetical software offered by BOLD for these clients. It could then access a specific BOLD node based on the higher taxon that needs to be identified.



This view would conform with the notion of the [Conservation Commons](#) or the emerging Biodiversity Commons (Moritz 2002), that also advocate the open access to biodiversity data. Collaborative efforts need not necessarily lack a rigorous structure (cf. Godfray 2002).

**BOLD should closely monitor developments of biodiversity portals (especially GBIF) in order to become the standard link between biodiversity and molecular information.** Thus, it should try not to just bring in data from external sources into its own architecture (see Fig. 19 with query links to ITIS or the phylogenetic Tree of Life project, [TOL](#)), but also to bring its content into external architectures.

Figure 19: BOLD queries to NCBI, ITIS and TOL



Section of a page from the BOLD Taxonomy Browser showing links (red box) programmed to query GenBank, ITIS, TOL and NCBI's Taxonomy database for the respective taxon (see text). Source:

[http://www.boldsystems.org/views/taxbrowser\\_root.php](http://www.boldsystems.org/views/taxbrowser_root.php)

### 3 Conclusion

The increasing application of molecular methods to taxonomy is as much a consequence of new technological possibilities as a systematic effort to compensate the 'taxonomic impediment', especially through automated species identifications. In the light of massive accumulation of data in molecular taxonomy, especially DNA barcoding, I focused on data management questions for these resources. For this purpose, I analyzed the NCBI [GenBank](#) database and the Barcode of Life Data System ([BOLD](#)). Both of these databases are needed, as they have different aims and advantages. BOLD provides specialized and differentiated access for the taxonomy and biodiversity community while GenBank offers a centralized gateway and an 'archive' for a broader range of users. Neither BOLD nor GenBank are currently suited as platforms for routine species description, but the special GenBank 'BARCODE' records and especially BOLD entries are mostly suited to current species identification demands, for which BOLD also provides an interface. However, more metadata categories should be considered. In the near future, BOLD will have to develop a concept how to deal with samples that are not associated with specimen information. During the annotation of records, BOLD and especially GenBank should make more extensive use of the many available means to standardize structure and content of their data. Thereby they would increase interoperability with other databases to facilitate data exchange and information retrieval. BOLD will have to expand its search options and will have to focus on additional sequence retrieval algorithms. Regarding the unique naming system by which the analyzed databases address their records, BOLD needs to accommodate information on record versioning. Compared to GenBank, BOLD's strength lies in keeping records both for the sequence and for the specimen. The databases will sooner or later have to adopt the globally unique naming system procured by Life Science Identifiers, both for record IDs but also for other data. Although not manually curated, barcoding data in GenBank and BOLD have a higher likelihood to be accurate than regular GenBank records. In BOLD this is thanks to how the site is conceived. Most of the increased accuracy is related to how taxonomy is carried out. Also, database errors stand a better chance to be corrected. Overall, preservation chances for DNA taxonomy information are good, since data volume and format homogeneity make the selection of data unnecessary. Another positive aspect is that (currently), no digital rights management applies. However, a problem arises from the fact that the data (easy to preserve by themselves) reside in databases and, to obtain their full meaning, have to be preserved within these structures. BOLD and especially GenBank should strive to fulfill the necessary criteria to become 'trusted repositories'. Therefore, a more stable mode of funding would be preferable (although probably at different levels for the two databases in question). It is likely that soon most taxonomic and biodiversity information will coalesce through integration of a multitude of different data sources. BOLD should closely

monitor developments of biodiversity portals (especially GBIF) in order to become the standard link between biodiversity and molecular information. If addressing current deficits and if adapting to future needs, it is likely that BOLD will become not only this, but also the standard access portal for molecular taxonomy – or, if expanding its infrastructure, maybe even for taxonomy in general. At least, together with special GenBank records, it will be an important cornerstone in keeping taxonomy coherent (cf. Polaszek & Wilson 2005). Careful management of molecular taxonomy data and sharing these data globally is of paramount importance for all of biology and beyond. More than 20 years ago, C. Pabo (1987) already recognized the central, propelling role of information in biology: "Unlike physics, which moved to an era of 'big science' because of the costs of equipment for high energy research, it may be information that drives molecular biology [and from today's perspective also biodiversity] into big science, and leads to a cooperative style of research".

## Acknowledgements

I want to thank the following people for their incentives, answers or time: N. Astrin, M. Seadle, W. Wägele, S. Ratnasingham, B. Misof, J. Dambach, A. Lesk, B. Robson, and several members of the NCBI/GenBank staff.

## References

- Agerer, R., Ammirati, J., Blanz, P., Courtecuisse, R., Desjardin, D. E., Gams, W., Hallenberg, N., Halling, R., Hawksworth, D. L., Horak, E., Korf, R. P., Mueller, G. M., Oberwinkler, F., Rambold, G., Summerbell, R. C., Triebel, D. & Watling, R. (2000). Always deposit vouchers. *Mycological Research* **104**, 642-644.
- Agosti, D. & Johnson, N. F. (2002). Taxonomists need better access to published data. *Nature* **417**(6886), 222.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**(3), 403-410.
- Ashburner, M. & Goodman, N. (1997). Informatics - genome and genetic databases. *Current Opinion in Genetics & Development* **7**(6), 750-756.
- Bard, J. B. L. & Rhee, S. Y. (2004). Ontologies in biology: Design, applications and future challenges. *Nature Reviews Genetics* **5**(3), 213-222.
- Barker, W. C. & Wu, C. H. (2005). Annotation of protein sequences. In: *Database annotation in molecular biology*. Ed(s): A. Lesk. pp. 131-148. Wiley & Sons: Chichester, West Sussex.
- Beach, J. H., Pramanik, S. & Beaman, J. H. (1993). Hierarchic taxonomic databases. In: *Advances in computer methods for systematic biology: Artificial intelligence, databases, computer vision*. Ed(s): R. Fortuner. pp. 241-256. Johns Hopkins University Press: Baltimore, MD.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research* **36**, D25-D30.



- Berendsen, H. J. (2003). Inter-union bioinformatics group report. *Acta Crystallographica Section D Biological Crystallography* **59**(4), 777-782.
- Berendsohn, W. G. (1995). The concept of potential taxa in databases. *Taxon* **44**(2), 207-212.
- Bisby, F. A. (2000). The quiet revolution: Biodiversity informatics and the Internet. *Science* **289**(5488), 2309-2312.
- Blaxter, M. (2003). Molecular systematics: Counting angels with DNA. *Nature* **421**(6919), 122-124.
- Blazej, R. G., Kumaresan, P. & Mathies, R. A. (2006). Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **103**(19), 7240-7245.
- BOLD - Barcode of Life Data Systems. URL (last accessed May 2008): <http://www.barcodinglife.org/>
- Borghoff, U. M., Rödig, P., Scheffczyk, J. & Schmitz, L. (2003). *Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente*. dpunkt: Heidelberg.
- Borghoff, U. M., Rödig, P., Scheffczyk, J. & Schmitz, L. (2005). *Langzeitarchivierung. Informatik Spektrum* **28**(6), 489-492.
- Brazma, A., Krestyaninova, M. & Sarkans, U. (2006). Standards for systems biology. *Nature Reviews Genetics* **7**(8), 593-605.
- Bridge, P. D., Roberts, P. J., Spooner, B. M. & Panchal, G. (2003). On the unreliability of published DNA sequences. *New Phytologist* **160**(1), 43-48.
- Brooke, M. D. (2000). Why museums matter. *Trends in Ecology & Evolution* **15**(4), 136-137.
- Brusic, V., Zeleznikow, J. & Petrovsky, N. (2000). Molecular immunology databases and data repositories. *Journal of Immunological Methods* **238**(1-2), 17-28.
- Canhos, V. P., Souza, S., Giovanni, R. & Canhos, D. A. L. (2004). Global biodiversity informatics: Setting the scene for a “new world” of ecological modeling. *Biodiversity Informatics* **1**, 1-13.
- Chase, M. W., Cowan, R. S., Hollingsworth, P. M., Van Den Berg, C., Madrinan, S., Petersen, G., Seberg, O., Jorgensen, T., Cameron, K. M., Carine, M., Pedersen, N., Hedderson, T. A. J., Conrad, F., Salazar, G. A., Richardson, J. E., Hollingsworth, M. L., Barraclough, T. G., Kelly, L. & Wilkinson, M. (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon* **56**(2), 295-299.
- Chu, K. H., Li, C. P. & Qi, J. (2006). Ribosomal RNA as molecular barcodes: A simple correlation analysis without sequence alignment. *Bioinformatics* **22**(14), 1690-1701.
- Clark, T., Martin, S. & Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics* **5**(1), 59-70.
- Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., Bhattacharyya, S., Bonfield, J., Bower, L., Browne, P., Castro, M., Cox, T., Demiralp, F., Eberhardt, R., Faruque, N., Hoad, G., Jang, M., Kulikova, T., Labarga, A., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Plaister, S., Robinson, S., Sobhany, S., Vaughan, R., Wu, D., Zhu, W. M., Apweiler, R., Hubbard, T. & Birney, E. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl trace archive and the embl nucleotide sequence database. *Nucleic Acids Research* **36**, D5-D12.
- Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M. & Schilling, L. M. (2003). Going, going, gone: Lost Internet references. *Science* **302**(5646), 787-788.
- De Solla Price, D. J. (1963). *Little Science, Big Science*. Columbia University Press: New York, NY.
- EDIT - European Distributed Institute of Taxonomy. (2007). *Report to the board of directors*. Conference: Taxonomy in Europe in the 21st century, Oxford. URL (accessed April 2008): <http://ww2.bgbm.org/EditDocumentRepository/Taxonomy21report.pdf>
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**(9), 755-763.
- Edwards, J. L., Lane, M. A. & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* **289**(5488), 2312-2314.

- Ekrem, T., Willassen, E. & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution* **43**(2), 530-542.
- Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Research* **8**(3), 186-194.
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Research* **8**(3), 175-185.
- Fitzhugh, K. (2006). DNA barcoding: An instance of technology-driven science? *BioScience* **56**(6), 462-463.
- Forster, R. (2003). To err is human. *Annals of Human Genetics* **67**, 2-4.
- Frishman, D., Heumann, K., Lesk, A. & Mewes, H. W. (1998). Comprehensive, comprehensible, distributed and intelligent databases: Current status. *Bioinformatics* **14**(7), 551-561.
- Galperin, M. Y. (2008). The molecular biology database collection: 2008 update. *Nucl. Acids Res.* **36 Supplement 1**, D2-4.
- Garnier, J. & Berendsen, H. J. (2002). International unions concerned about biodata. *Nature* **419**(6909), 777.
- Garrity, G. M. & Lilburn, T. G. (2005). Self-organizing and self-correcting classifications of biological data. *Bioinformatics* **21**(10), 2309-2314.
- GenBank. URL (last accessed May 2008): <http://www.ncbi.nlm.nih.gov/Genbank/>
- Gewin, V. (2002). All living things, online. *Nature* **418**(6896), 362-363.
- Godfray, H. C. (2002). Challenges for taxonomy. *Nature* **417**(6884), 17-19.
- Graham, M. & Kennedy, J. (2007). Visual exploration of alternative taxonomies through concepts. *Ecological Informatics* **2**(3), 248-261.
- Greenstone, M. H., Rowley, D. L., Heimbach, U., Lundgren, J. G., Pfannenstiel, R. S. & Rehner, S. A. (2005). Barcoding generalist predators by polymerase chain reaction: Carabids and spiders. *Mol Ecol* **14**(10), 3247-3266.
- Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature* **434**(7037), 1067.
- Hajibabaei, M., Dewaard, J. R., Ivanova, N. V., Ratnasingham, S., Dooh, R. T., Kirk, S. L., Mackie, P. M. & Hebert, P. D. (2005). Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences* **360**(1462), 1959-1967.
- Hanner, R. (2005). *Proposed standards for BARCODE records in INSDC (BRIs)*. URL (accessed April 2008): [http://www.barcoding.si.edu/PDF/DWG\\_data\\_standards-Final.pdf](http://www.barcoding.si.edu/PDF/DWG_data_standards-Final.pdf)
- Hansen, A. (2004). *Bioinformatik*. Birkhäuser: Basel.
- Harris, D. J. (2003). Can you bank on GenBank? *Trends in Ecology & Evolution* **18**(7), 317-319.
- Hawksworth, D. L. (1995). *Global biodiversity assessment*. Cambridge, Cambridge University Press.
- Hebert, P. D., Cywinska, A., Ball, S. L. & Dewaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **270**(1512), 313-321.
- Hebert, P. D., Ratnasingham, S. & Dewaard, J. R. (2003b). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **270 Supplement 1**, S96-99.
- Hillis, D. M., Mable, B. K., Larson, A., Davis, S. K. & Zimmer, E. A. (1996). Nucleic acids IV: Sequencing and cloning. In: *Molecular systematics*. Ed(s): D. M. Hillis, C. Moritz & B. K. Mable. Sinauer: Sunderland, MA.
- Hine, C. (2008). *Systematics as cyberscience: Computers, change, and continuity in science*. MIT Press: Cambridge, MA.
- Huber, J. T. (1998). The importance of voucher specimens, with practical guidelines for preserving specimens of the major invertebrate phyla for identification. *Journal of Natural History* **32**(3), 367-385.

- iBOL - International Barcode of Life. URL (accessed April 2008): <http://www.dnabarcoding.org/iw/presentations.htm>
- ISI Web of Knowledge. URL (accessed March 2008): <http://isiknowledge.com/>
- Kahle, B. (1997). Preserving the Internet. *Scientific American* **276**(3), 72-74.
- Kelly, R. P., Sarkar, I. N., Eernisse, D. J. & Desalle, R. (2007). DNA barcoding using chitons (genus *Mopalia*). *Molecular Ecology Notes* **7**(2), 177-183.
- Kennedy, J. B., Kukla, R. & Paterson, T. (2005). Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. *Data Integration in the Life Sciences, Proceedings* **3615**, 80-95.
- Knapp, S., Lamas, G., Lughadha, E. N. & Novarino, G. (2004). Stability or stasis in the names of organisms: The evolving codes of nomenclature. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences* **359**(1444), 611-622.
- Koehler, W. (2002). Web page change and persistence - a four-year longitudinal study. *Journal of the American Society for Information Science and Technology* **53**(2), 162-171.
- Kress, W. J. & Erickson, D. L. (2008). DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences* **105**(8), 2761-2762.
- Krichevsky, M. I. (2005). Taxonomy: A moving target for sequence data. In: *Database annotation in molecular biology*. Ed(s): A. Lesk. pp. 101-112. Wiley & Sons: Chichester, West Sussex.
- Kumar, S. & Dudley, J. (2007). Bioinformatics software for biologists in the genomics era. *Bioinformatics* **23**(14), 1713-1717.
- Linné, C. (1758). *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Salvius: Stockholm.
- Lipscomb, D., Platnick, N. & Wheeler, Q. (2003). The intellectual content of taxonomy: A comment on DNA taxonomy. *Trends in Ecology & Evolution* **18**(2), 65-66.
- Liu, Z. Z., Lozupone, C., Hamady, M., Bushman, F. D. & Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research* **35**(18), e120.
- Lughadha, E. N. (2004). Towards a working list of all known plant species. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences* **359**(1444), 681-687.
- Lyman, P. & Varian, H. R. (2003). *How much information?* URL (accessed April 2008): <http://www.sims.berkeley.edu/how-much-info-2003>
- Mace, G. M. (2004). The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences* **359**(1444), 711-719.
- Machlis, G. E. (1992). The contribution of sociology to biodiversity research and management. *Biological Conservation* **62**(3), 161-170.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. T., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., Mckenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P. G., Begley, R. F. & Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057), 376-380.
- May, R. M. (1990). Taxonomy as destiny. *Nature* **347**, 129-130.
- May, R. M. & Lawton, J. H. (1995). *Extinction rates*, Oxford University Press.
- Min, X. J. & Hickey, D. A. (2007). Assessing the effect of varying sequence length on DNA barcoding of fungi. *Molecular Ecology Notes* **7**(3), 365-373.
- Minelli, A. (2003). The status of taxonomic literature. *Trends in Ecology & Evolution* **18**(2), 75-76.

- Moritz, T. (2002). Building the Biodiversity Commons. *D-Lib Magazine* **8**(6).  
URL: <http://www.dlib.org/dlib/june02/moritz/06moritz.html>.
- Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis*. Cold Spring Harbor, New York, NY, Cold Spring Harbor Laboratory Press.
- NCBI-*GenBank Flat File Release*: cf. version number and accession date in the text – current versions can be accessed under the URL <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>
- NCBI - National Center for Biotechnology Information. *NCBI Taxonomy*. URL (accessed Feb. 2008): <http://www.ncbi.nlm.nih.gov/Taxonomy/>
- NCBI - National Center for Biotechnology Information. *Barcode Submission Tool*. URL (accessed April 2008): <http://www.ncbi.nlm.nih.gov/WebSub/?tool=barcode>
- NC-IUB - Nomenclature Committee of the International Union of Biochemistry (1985). Nomenclature for incompletely specified bases in nucleic-acid sequences - Recommendations 1984. *Biochemical Journal* **229**(2), 281-286.
- Neale, S. H., Pullan, M. R. & Watson, M. F. (2007). Online biodiversity resources – principles for usability. *Biodiversity Informatics* **4**, 27-36.
- nestor - Kompetenznetzwerk Langzeitarchivierung (2006). *Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland*. URL (accessed May 2008): <http://www.langzeitarchivierung.de/downloads/memo2006.pdf>
- NLM - National Library of Medicine. *The NCBI Handbook*. URL (accessed Mar. 2008): <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook>
- Pabo, C. O. (1987). New generation databases for molecular biology. *Nature* **327**, 467.
- Page, R. (2005). A taxonomic search engine: Federating taxonomic databases using web services. *BMC Bioinformatics* **6**(1), 48.
- Page, R. D. M. (2006). Taxonomic names, metadata, and the semantic web. *Biodiversity Informatics* **3**, 1-15.
- Parr, C. S. & Cummings, M. P. (2005). Data sharing in ecology and evolution. *Trends in Ecology & Evolution* **20**(7), 362-363.
- Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. (1995). The future of biodiversity. *Science* **269**(5222), 347-350.
- Polaszek, A. (2005). A universal register for animal names. *Nature* **437**(7058), 477.
- Polaszek, A. & Wilson, E. O. (2005). Sense and stability in animal names. *Trends in Ecology & Evolution* **20**(8), 421-422.
- Rach, J., Desalle, R., Sarkar, I. N., Schierwater, B. & Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B: Biological Sciences* **275**(1632), 237-247.
- Ratnasingham, S. & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System. *Molecular Ecology Notes* **7**(3), 355-364.
- RLG (Research Library Group) - RLG/OCLC Working Group on Digital Archive Attributes (2002). Trusted digital repositories: Attributes and responsibilities. URL (accessed May 2008): <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>
- Ronquist, F. & Gärdenfors, U. (2003). Taxonomy and biodiversity inventories: Time to deliver. *Trends in Ecology & Evolution* **18**(6), 269-270.
- Ross, H. A. & Murugan, S. (2006). Using phylogenetic analyses and reference datasets to validate the species identities of cetacean sequences in GenBank. *Molecular Phylogenetics and Evolution* **40**(3), 866-871.
- Ruedas, L. A., Salazar-Bravo, J., Dragoo, J. W. & Yates, T. L. (2000). The importance of being earnest: What, if anything, constitutes a "specimen examined?" *Molecular Phylogenetics and Evolution* **17**(1), 129-132.
- Saitou, N. & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**(4), 406-425.
- Sarkar, I. N., Planet, P. J., Bael, T. E., Stanley, S. E., Siddall, M., Desalle, R. & Figurski, D. H. (2002). Characteristic attributes in cancer microarrays. *Journal of Biomedical Informatics* **35**(2), 111-122.

- Schulze-Kremer, S. (2002). Ontologies for molecular biology and bioinformatics. *In Silico Biology* **2**(17).
- Schwens, U. & Liegmann, H. (2004). Langzeitarchivierung digitaler Ressourcen. In: *Grundlagen der praktischen Information und Dokumentation*. Ed(s): K. Laisiepen, E. Lutterbeck & K.-H. Meyer-Uhlenried. pp. 567-570. Saur: Munich.
- Seberg, O., Humphries, C. J., Knapp, S., Stevenson, D. W., Petersen, G., Scharff, N. & Andersen, N. M. (2003). Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends in Ecology & Evolution* **18**(2), 63-65.
- Shaffer, C. (2007). Next-generation sequencing outpaces expectations. *Nature Biotechnology* **25**(2), 149-149.
- Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W. & Gillevet, P. M. (1994). The genetic data environment: An expandable GUI for multiple sequence analysis. *Computer Applications in the Biosciences* **10**(6), 671-675.
- Smith, T. F. (2002). The challenges facing genomic informatics. In: *Current topics in computational molecular biology*. Ed(s): T. Jiang, Y. Xu & M. Q. Zhang. pp. 3-8. MIT Press: Cambridge, MA.
- Soberon, J. (1999). Linking biodiversity information sources. *Trends in Ecology & Evolution* **14**(7), 291-291.
- Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T. & Tateno, Y. (2008). DDBJ with new system and face. *Nucleic Acids Research* **36**, D22-D24.
- Swaminathan, G. J., Tate, J., Newman, R., Hussain, A., Ionides, J., Henrick, K. & Velankar, S. (2005). Issues in the annotation of protein structures. In: *Database annotation in molecular biology*. Ed(s): A. Lesk. pp. 149-165. Wiley & Sons: Chichester, West Sussex.
- Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology & Evolution* **18**(2), 70-74.
- Thibodeau, K. (2002). *Overview of technological approaches to digital preservation and challenges in coming years*. URL (accessed May 2008): <http://www.clir.org/pubs/reports/pub107/thibodeau.html>
- Thiele, K. & Yeates, D. (2002). Tension arises from duality at the heart of taxonomy. *Nature* **419**(6905), 337.
- UNESCO - United Nations Educational, Scientific and Cultural Organization (2003). *Guidelines for the preservation of digital heritage*. URL (accessed Jan. 2008): <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>
- Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. & Yaschenko, E. (2008). Database resources of the national center for biotechnology information. *Nucleic Acids Research* **36**, D13-D21.
- Wilson, E. O. (2003). The encyclopedia of life. *Trends in Ecology & Evolution* **18**(2), 77-80.
- Wilson, E. O. (2004). Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences* **359**(1444), 739-739.
- Wouters, P. & Beaulieu, A. (2002). *Quality control of data in data-sharing practices and regulations*. Conference: CODATA 2002, Frontiers of Scientific and Technical Data. Montréal, Canada. URL (accessed April 2008): <http://www.codata.org/codata02/abs-5theme.html>
- Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press: New York, NY.

## Appendix: Internet resources

ABCD, Access to Biological Collections Data: <http://www.tdwg.org/activities/abcd/>  
Algaebase: <http://www.algaebase.org/>  
All Birds Barcoding Initiative: <http://www.barcodingbirds.org/>  
All Species Foundation: <http://www.all-species.org/>  
All-Leps Barcode of Life: <http://www.lepbarcoding.org/>  
Antbase: <http://antbase.org/>  
ASN.1, Abstract Syntax Notation number One: <http://asn1.elibel.tm.fr/>  
Bacterial taxonomy register: <http://www.bacterio.cict.fr/>  
Barcode of Life Initiative: <http://www.dnabarcodes.org/>  
BHL, Biodiversity Heritage Library: <http://www.biodiversitylibrary.org/>  
BioCASE: <http://www.biocase.org/products/protocols/>  
BLAST, Basic Local Alignment Search Tool: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>  
BOLD alternative markers: [http://www.barcoding.si.edu/PDF/Guidelines for non-CO1 selection FINAL.pdf](http://www.barcoding.si.edu/PDF/Guidelines%20for%20non-CO1%20selection%20FINAL.pdf)  
BOLD Identification Engine: <http://www.boldsystems.org/views/idrequest.php>  
BOLD Taxonomy Browser: [http://www.barcodinglife.org/views/taxbrowser\\_root.php](http://www.barcodinglife.org/views/taxbrowser_root.php)  
BOLD, Barcode of Life Data System: <http://www.barcodinglife.org/>  
Canadian Barcode of Life Network: <http://www.bolnet.ca/campaigns.php>  
CBD, Convention on Biological Diversity: <http://www.cbd.int/>  
CBOL, Consortium for the Barcode of Life: <http://www.barcoding.si.edu/>  
CCDB, Canadian Centre for DNA Barcoding: <http://www.dnabarcoding.ca/>  
Census of Marine Zooplankton: <http://www.emarz.org/>  
CODATA, Committee on Data for Science and Technology: <http://www.codata.org/>  
CoL, Catalogue of Life: <http://www.catalogueoflife.org/>  
Conservation Commons: <http://conservationcommons.org/>  
CORBA: <http://www.omg.org/technology/documents/formal/components.htm>  
Darwin Core: <http://www.tdwg.org/activities/darwincore/>  
DB2: <http://www-306.ibm.com/software/data/db2/>  
DCMI, Dublin Core Metadata Initiative: <http://dublincore.org/>  
DDBJ, DNA Data Bank of Japan: <http://www.ddbj.nig.ac.jp/>  
DiGIR, Distributed Generic Information Retrieval: <http://digir.sourceforge.net/>  
DOI, Digital Object Identifier: <http://www.doi.org/>  
Dublin Core: <http://dublincore.org/documents/dces/>  
EMBL, European Molecular Biology Laboratory: <http://www.embl.org/>  
EMBL-Bank: <http://www.ebi.ac.uk/embl/>  
Entrez Gene: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>  
Entrez: <http://www.ncbi.nlm.nih.gov/Database/>  
EOL, Encyclopedia of Life: <http://www.eol.org/>  
Firefox extension for LSIDs: <http://sourceforge.net/projects/lisids/>  
Fish Barcode of Life Initiative: <http://www.fishbol.org/>  
Fish Barcode of Life: <http://www.fishbol.org/>  
Florida Digital Archive: <http://www.fcla.edu/digitalArchive/>  
GBIF, Global Biodiversity Information Facility: <http://data.gbif.org/>  
GenBank source modifiers: <http://www.ncbi.nlm.nih.gov/Sequin/modifiers.html>  
GenBank: <http://www.ncbi.nlm.nih.gov/Genbank/>  
Gene Ontology: <http://www.geneontology.org/>  
Getty Trust: <http://www.getty.edu/>  
GML, Geography Markup Language: <http://www.opengeospatial.org/standards/gml>  
GNS, GEOnet Names Server: <http://earth-info.nga.mil/gns/html/>  
HUGO Nomenclature Committee: <http://www.genenames.org/>  
HUGO, Human Genome Organisation: <http://www.hugo-international.org/>

iBOL, International Barcode of Life Initiative: <http://www.dnabarcoding.org/>  
ICZN, International Commission on Zoological Nomenclature: <http://www.iczn.org/>  
Index Fungorum: <http://www.indexfungorum.org/Names/Names.asp>  
INSDC, International Nucleotide Sequence Database Collaboration: <http://www.insdc.org/>  
IPNI, International Plant Names Index: <http://www.ipni.org/>  
iSpecies: <http://darwin.zoology.gla.ac.uk/~rpage/ispecies/>  
ITIS, Integrated Taxonomic Information System: <http://www.itis.gov/>  
IUBS, International Union of Biological Sciences: <http://www.iubs.org/>  
LSID resolver: <http://lsid.tdwg.org/>  
LSIDs, Life Science Identifiers: <http://lsids.sourceforge.net/>  
Marine Barcode of Life: <http://www.marinebarcoding.org/>  
My NCBI: <http://www.ncbi.nlm.nih.gov/entrez/cubby.fcgi>  
Mycobank: <http://www.mycobank.org/>  
NCBI data mining tools: <http://www.ncbi.nlm.nih.gov/Tools/>  
NCBI taxonomy database: <http://www.ncbi.nlm.nih.gov/Taxonomy/>  
NCBI Trace Archive: <http://www.ncbi.nlm.nih.gov/Traces>  
NCBI, National Center for Biotechnology Information: <http://www.ncbi.nlm.nih.gov/>  
NIG, National Institute of Genetics: <http://www.nig.ac.jp/>  
OAIS, Open Archival Information System: <http://public.ccsds.org/publications/archive/650x0b1.pdf>  
OGC, Open Geospatial Consortium: <http://www.opengeospatial.org/>  
OMG, Object Management Group: <http://www.omg.org/>  
Polar Barcode of Life: <http://www.polarbarcoding.org/>  
PostgreSQL: <http://www.postgresql.org/>  
PRONOM file format registry: <http://www.nationalarchives.gov.uk/PRONOM/>  
PubMed: <http://www.ncbi.nlm.nih.gov/pubmed/>  
PURL, Persistent Uniform Resource Locator: <http://purl.org/>  
RefSeq: <http://www.ncbi.nlm.nih.gov/RefSeq/>  
sequence formats: <http://emboss.sourceforge.net/docs/themes/SequenceFormats.html>  
SOAP: <http://www.w3.org/TR/soap12-part1/>  
Species 2000: <http://www.sp2000.org/>  
Sponge Barcoding Project: <http://www.spongebarcoding.org/>  
SwissProt: <http://www.expasy.org/sprot/>  
TAPIR: <http://www.tdwg.org/activities/tapir/>  
TCS, Taxonomic Concept Transfer Schema: <http://www.tdwg.org/standards/117/>  
TDWG, Biodiversity Information Standards group: <http://www.tdwg.org/>  
TGN, Thesaurus of Geographic Names: [http://www.getty.edu/research/conducting\\_research/vocabularies/tgn/](http://www.getty.edu/research/conducting_research/vocabularies/tgn/)  
Thesaurus of the Zoological Record: <http://scientific.thomsonreuters.com/support/products/zr/thesaurus/>  
TOL, Tree of Life: <http://www.tolweb.org/tree/>  
TPA, NCBI Third Party Annotations: <http://www.ncbi.nlm.nih.gov/Genbank/TPA.html>  
UniProt: <http://www.pir.uniprot.org/>  
Wikispecies: <http://species.wikimedia.org/>  
WIPO, World Intellectual Property Organization: <http://www.wipo.int/>  
XML, Extensible Markup Language: <http://www.w3.org/XML/>  
Z39.50: <http://www.loc.gov/z3950/agency/>  
ZooBank: <http://www.iczn.org/ZooBank.html>  
Zoological Record: <http://thomsonscientific.com/products/zr/>