URL: http://www.iupac.org/symposia/proceedings/phuket97/oliver.html

A Virtual Catalogue of Invertebrate Life on Earth

Ian Oliver

Key Centre for Biodiversity and Bioresources, School of Biological Sciences, Macquarie University, Sydney, NSW, 2109 Australia

Abstract: In signing the Convention on Biological Diversity, more than 160 countries pledged to produce inventories of their biodiversity - workable catalogues of the variety of life. Some countries have already passed legislation to include this biological information in land-use planning processes. Invertebrate diversity remains largely unknown despite their presence in all habitats and critical role in the maintenance of ecosystem health and ultimately, human well-being. Recent estimates suggest that there may be as many as 30 million species of arthropods alone, yet taxonomists have formally identified less than one million. Consequently, species inventories that document the numbers and identities of invertebrates in space and time are difficult to produce and the catalogue of the Earth's invertebrate biodiversity is very incomplete.

New and exciting methodological and technological solutions to this global problem are now emerging, for example: the use of local non-specialist taxonomic workers to increase rates of specimen collection and processing; morphospecies in place of formal species names (Latin binomials) and the most recent of bioinformatics technologies. The use of bioinformatics (application of information technology to the solution of biological problems) includes: the addition of unique micro-barcodes to samples and specimens to remove the need for hand written labels and increase the speed of data entry and retrieval; and sophisticated biodiversity database management systems for the capture, storage, manipulation and retrieval of biodiversity data. These data may include taxonomic and ecological information or digital images of species which can be used as virtual voucher specimens in the process of sorting and identification. This paper discusses these recent advances and explores the potential of bioinformatics to provide a digital image database or virtual catalogue of the Earth's invertebrate biodiversity.

THE PAST

Size of the invertebrate fauna

The Linnean system of classification is the preferred approach for separating and cataloguing species level biodiversity. Once classified, each different type of organism receives a unique

*Invited lecture presented at the International Conference on Biodiversity and Bioresources: Conservation and Utilization, 23–27 November 1997, Phuket, Thailand. Other presentations are published in *Pure Appl. Chem.*, Vol. 70, No. 11, 1998.

label in the form of a Latin binomial representing it as a unique biological species. Although this system is not without its limitations (e.g. microorganisms which freely exchange DNA between species), for animals and plants it is generally transferable across regions and between organisms. During the 250 years since the Linnean system was introduced, approximately 1.7 million species of animals, plants and microorganisms have been described and named by taxonomists. Approximately 3% of these species are vertebrates, 16% are plants and 74% are invertebrates, the remaining 7% comprise the microorganisms (ref. 1). Invertebrates, therefore, constitute the bulk of described biodiversity, a group which is dominated by the arthropods (65% of all described species) and which includes the most notable taxon, the beetles, which represent 24% of all described species.

Comparison of these figures to recent estimates of total (undescribed and unknown) biodiversity illustrates how little we know of the extent and distribution of invertebrates. Until recently, the total number of species on Earth was believed to lie in the range of 3 to 5 million (ref. 2). Following a study of tropical rainforest canopy beetles, Terry Erwin (ref. 3) increased this estimate by an order of magnitude to 30 to 50 million species of arthropods alone. Other estimates have gone as high as 80 million species of arthropods (ref. 4) and still no consensus has been reached (refs. 5-10). In short, most non-microbial biodiversity is comprised of invertebrates, a large proportion of that by arthropods, most of which have not been formally identified nor even collected.

Traditional approaches to cataloguing invertebrate biodiversity

The description and naming of species (taxonomy) has been the realm of the specialist taxonomist for more than two centuries. For species-rich groups such as the invertebrates, identification of a species new to science is not an easy task and is normally included as part of a monograph which often concentrates on many similar species from a particular area. Monographs also seek to organise large numbers of species by classifying them in a hierarchical manner which aims to reflect their evolutionary history (systematics). For example, single genera should contain species which share similar morphological characteristics and therefore a common recent ancestor. These phylogenetic classifications are valuable because they allow researchers to make generalisations about poorly known species based on species with similar phylogenies whose biologies are better understood (ref. 11). However, this approach is slow and faces many constraints (see Box 1) which means that if the description of new species, using traditional methods, continues at the same rate with the same number of taxonomists as today, the cataloguing of global biodiversity will take several thousand years to complete (refs. 12, 13). Given the extent of invertebrate biodiversity new more rapid approaches are required because the parties to the Convention on Biological Diversity will have great difficulties honouring their international obligations if traditional taxonomic approaches are their only means.

Box 1. Constraints and problems faced by traditional taxonomic approaches

- The number of professional taxonomists continues to decline, as does funding for taxonomic research and the numbers of university graduates who specialise in taxonomy (refs. 11, 12, 14).
- Levels of synonymy (the allocation of the different Latin binomials to the same biological species) is commonly around 20% and may be as high as 50% of the total number of described species in some invertebrate groups (ref. 15).
- Identification of invertebrate specimens to species by non-specialists is often impossible due to an absence of scientific keys. Most of the available keys use very specialised terminology which are impossible to follow without significant training.
- The geographical distribution of taxonomists is poorly matched to the species richness of biogeographical regions. For example, 80% of the worlds insect taxonomists are located in North America and Europe and only 7% in the neo-tropical and Ethiopian regions (ref. 14).
- The distribution of taxonomists among taxa poorly matches the potential numbers of those taxa. For example, there are an estimated three unrecorded tetrapod species, 34 unrecorded fish species and 400 unrecorded invertebrate species for each professional taxonomist working with these particular groups (ref. 14).
- The space required to store representatives of each invertebrate species is considerable. For example, representatives of only 500,000 insect species currently fill six floors of London's Natural History Museum (ref. 11). Representatives of a potential ten million insect species might therefore require a building taller than the Empire State Building (top floor height = 391m above ground level).

THE PRESENT

Attempts to conserve and utilise invertebrate biodiversity are hampered by our lack of knowledge of the extent, distribution and biology of species and traditional approaches to cataloguing invertebrate biodiversity can not provide this information quickly. Many biodiversity researchers and taxonomists are aware of the constraints of traditional approaches and have developed and tested a number of new methodologies and are increasingly utilising bioinformatics. These new approaches to the collection of invertebrate biodiversity data, its storage and management, the identification of species and the dissemination of information provide the opportunity to meet international obligations.

Data collection, storage and management

The first stage in increasing understanding of biodiversity is to collect specimens and species at a faster rate. Traditionally, the collection of species was the domain of the specialist taxonomist. More recently a new profession has been created, the parataxonomist or biodiversity technician. They may be people with little or no biological training or university graduates with little or no formal training in taxonomy. Training in parataxonomy instead covers collection and preparation of invertebrates, an introduction to species identification and conveys an understanding of the

importance of biodiversity. The main role of the parataxonomist is specimen collection, however, they are becoming increasingly involved with the sorting and identification of specimens. In Costa Rica parataxonomists are collecting more than 100,000 insect specimens a month and the 100 to 200 parataxonomists that will eventually be trained are expected to put 95% of the nations arthropod diversity (estimated to include 300,000 insect species) into collections within 10 years (refs. 16-18). Pioneered in Costa Rica, this approach and the new workforce it has created continues to grow throughout the world (refs. 19-21).

Recent increases in the rates of specimen collection has resulted in the generation of huge numbers of specimens and data records. For example, fogging the canopies of tropical forest trees with a pyrethroid insecticide may return thousands of species and tens of thousands of individuals (refs. 3, 22). At least each species, but more often several specimens of each species, must be labelled so that the associated taxonomic, ecological and accession information can be "tagged" to that specimen. The use of bioinformatics including commercial barcodes and the development of new biodiversity database management systems, enables biodiversity workers to keep track of this rapidly accumulating data (see Box 2).

Box 2. Barcodes and biodiversity database management systems

The addition of unique barcodes to each sample or specimen is quick, removes the need for hand written labels and increases the speed at which data is entered into and retrieved from biodiversity databases. One American taxonomist estimated that the use of barcodes would have saved him a quarter of a million keystrokes in one year (ref. 23). The use of barcodes also improves data quality by removing transcription and typing errors. Invertebrate biodiversity workers are embracing new barcode technology and have been integral to the miniaturisation of labels. "Code-49" labels are the smallest and use "dense code" which enables the incorporation of more than 100 alpha-numeric characters into an area 8x13mm (ref. 24).

Efficient storage and management of the data linked to each unique barcode is critical and has resulted in the development of several biodiversity database management systems such as

- ALICE (http://dspace.dial.pipex.com/alice)
- BIOTA (http://viceroy.eeb.uconn.edu/biota)
- FLORIN (http://www.florin.ru/florin)
- MUSE (http://biodiversity.bio.uno.edu/muse),
- PANDORA(http://www.rbge.org.uk/pankey.html) and
- Linnaeus II (http://wwweti.bio.uva.nl/demo/folnvgtr/hnvgtr.html).

At the Key Centre for Biodiversity and Bioresources we are using Biota for data management. Biota is a specimen-based, biological data and collections management system specifically designed for the ecological and taxonomic requirements of biodiversity data. The application is a fully relational data base system into which you enter your own specimen based biodiversity data. It has been designed from the outset to take advantage of barcode technology and the input compression and storage of species images. It is affordable as a single user application (US$125)

and available for Macintosh and Windows platforms (multi-platform client/server versions are also available).

Specimen identification

Biodiversity data is of little use in the absence of specimen identification but identification to Latin binomials is impossible for most invertebrate taxa because descriptions have yet to be written. One methodological solution to this taxonomic impediment is to (temporarily) abandon the use of the formal species name (Latin binomial) and identify taxa to morphospecies only. Morphospecies are determined using the external characters of specimens, often by non-specialist personnel, that is, parataxonomists or biodiversity technicians, and often approximate formal identifications made by specialist taxonomists (refs. 21, 25). This approach, which is becoming more widespread, applies a unique code number to species within a local framework (refs. 26, 27). For example, a biodiversity study may be interested in comparing the overlap of arthropod species sampled from the canopies of several different tropical tree species in close proximity. The biodiversity researcher allocates unique species codes to each morphospecies as the samples are processed and these codes are used in place of formal species names in subsequent analyses (refs. 26, 28).

The application of bioinformatics to the process of specimen identification has taken several paths and resulted in the development of some very useful multi-access keys, expert systems and approaches which use hypertext or neural networks (see Box 3). All require large amounts of specialist input and are often restricted to a particular taxon and named species only. Despite their limitations these systems do demonstrate the potential offered by bioinformatics, for example, one person cannot easily comprehend 3000 termite species whereas a computer has no trouble storing, searching and retrieving this information. Global biodiversity catalogues will only be possible with these sorts of computer intensive approaches.

Box 3. Application of bioinformatics to species identification.

Hypertext

The application of hypertext to species identification has been in the computerised representation of dichotomous and polytomous keys. As with hardcopy keys, the user is presented with two or more character states and must choose the most appropriate for the specimen in question. This selection then determines the next set of character states to choose between, and so-on until a final determination is made. The advantages of a hypertext based system are that the user can quickly retrace their tracks, a process that is difficult using conventional keys, and graphics and sound or other multi-media may be incorporated (ref. 29).

Multi-access keys

These identification guides use a data matrix of species by character combinations. Each cell in the matrix contains a value representing the state of the character for each species represented in the matrix. The important difference between multi-access keys and the traditional dichotomous key is that the user decides in which order character states are entered rather than following a prescribed sequence as presented by a dichotomous key. The advantage of this system is that the most obvious characters are entered first minimising the chance of incorrectly identifying

uncertain characters. However, if too few characters are entered the key may not be able to make a unique identification and may present several possibilities (ref. 29).

Expert systems

Expert systems have been defined as sophisticated computer programs that manipulate knowledge to solve problems efficiently and effectively in a narrow problem area and have been likened to an expert on the end of a phone. The system works under a question-answer framework with the computer (expert) asking questions of the investigator in order to help identify the specimen. Expert systems are different from multi-access keys because an expert system controls the problem-solving strategy and can be asked to provide explanations of its decisions and requests for information. It is also able to cope with uncertain or incomplete knowledge and still come to a conclusion or provide some advice even if the specimen does not exactly match any of the known species (ref. 29).

Neural networks

A neural network is a computer-intensive approach structured around layers of nodes which aim to parallel the processing of information by neurones in the brain. Each node in the input layer receives one or more pieces of information from the user which is summed then weighted and if the value exceeds a threshold value the result is automatically input into other nodes in the next layer. The process attempts to match the patterns represented by the input data to patterns representing known species. The user supplies the information on all the taxonomic characters at once and no information is provided as to how the identification was determined.

Neural networks depend on a period of supervised learning to match input patterns to known species. This requires a matrix of specimens by characters for each species rather than a species by character matrix. By this process the neural network learns the extent of both intra- and inter-specific variation. Training does require similar numbers of specimens for each species and the utility of the final application is dependent on the training set representing the full range of variation to be encountered (refs. 29, 30).

THE FUTURE

Building a virtual catalogue of invertebrate biodiversity using digital images - a view from the KCBB.

The limitations of past and present attempts to catalogue invertebrate biodiversity presented in this paper have led the Key Centre for Biodiversity and Bioresources to adopt the use of morphospecies in place of Latin binomials where necessary and parataxonomists or biodiversity technicians for routine sorting of specimens. Our aim is to employ digital image based identification of specimens, both at a local level (to morphospecies) and at a global level (to Latin binomials). The generation of high quality colour images of voucher specimens is a first step towards the development of a virtual catalogue of invertebrate biodiversity.

At the Key Centre, image capture is made using a scientific quality colour video camera on top of a stereo microscope. The camera contains 3 sensors which separately capture the red, blue and green bandwidths. The camera signal is digitised by a colour framegrabber installed

within a personal computer. Composing and focussing of the image can be done very quickly and on the computer monitor with an image resolution of 768x576 (PAL) pixels. Much higher resolutions are achievable with digital cameras although on screen composition and focussing is more difficult because these cameras take much longer to acquire an image. Digital cameras also tend to require more light and produce much larger file sizes.

Local level identification of specimens to morphospecies makes use of digital images in the following ways. When a new invertebrate morphospecies is encountered several images which capture the important taxonomic characters are acquired. These images are pasted into the biodiversity database management system *Biota* and linked to the appropriate taxonomic and ecological records. Each image represents about 1.5 megabytes of data but is compressed to about 0.2 megabytes before it is saved in the database. This process builds a database containing high quality images of all the morphospecies voucher specimens encountered. As new samples are processed the image database can be queried to determine whether specimens under the microscope are new morphospecies or have already been encountered and allocated a unique identifier. For example, the investigator may know that a new specimen may belong to the ant genus *Melophorus*, but its unique identification number may be unknown. A search can be made on the genus name and scrollable “thumb-nail” images of all morphospecies belonging to the genus *Melophorus* presented to the screen. Sub-selections of these can be made before individual images are requested in full size for comparison with the specimen of interest under the microscope. Identification is therefore achieved with reference to a virtual voucher specimen, a process which is quick, can simultaneously be undertaken by different investigators and removes the need for repeated handling of valuable voucher specimens. It does not however remove the need to keep and maintain the original samples and voucher specimens in good condition for later use by specialist taxonomists.

The second stage in creating a virtual catalogue of biodiversity is for these databases which include high quality graphics of morphospecies and their associated taxonomic and ecological information to go "on-line". This will facilitate global identification of specimens to Latin binomials through rapid access to virtual voucher material throughout the world. The databases will be able to be accessed and browsed by any machine connected to the world-wide-web. National networks such as the Australian Biological Research Network (ABREN, see http://lorenz.mur.csu.edu.au/abren/) are being established to manage the quality of the information provided to nodes on these networks. The Key Centre for Biodiversity and Bioresources is one of the first nodes on this network and its first contribution to it will be to provide browseable/searchable access to its virtual catalogues of invertebrate biodiversity (see Box 4).

Box 4. Possibilities presented by a virtual catalogue of invertebrate life on Earth

- On-line verification and identification of morphospecies by taxonomic specialists.
- Interrogation of other databases by non-specialists to assist in the identification of morphospecies.
- Reduced levels of synonymy because specialists undertaking revisions can rapidly accumulate virtual voucher specimens from anywhere in the world.
- Fewer losses of type specimens because specimens will not be required to sent around the world.
- Digital backups of type specimens held at multiple locations.
- The requirement for central repositories of voucher material will be reduced, therefore sharing the costs and space requirements amongst the broader scientific community.

ACKNOWLEDGMENTS

Thanks are extended to Annette Davison, George Tzotzos, Dale Oxender and the organising committee of the international conference "Biodiversity and Bioresources: Conservation and Utilisation" and to Mark Dangerfield, Andrew Beattie and Anthony Pik for comments on this manuscript. This is contribution number 252 to the Key Centre for Biodiversity and Bioresources, Macquarie University.

REFERENCES

1. B. Groombridge (ed). Global Biodiversity: Status of the Earth's living resources. Chapman and Hall, London (1992).
2. R. M. May. Nature 324, 514-515 (1986).
3. T. L. Erwin. The Coleop. Bull. 36, 74-75 (1982).
4. N. E. Stork. Biol. J. Linn. Soc. 35, 321-337 (1988).
5. R. M. May. Science 241, 1441-1449 (1988).
6. J. Adis. J. Trop. Ecol. 6, 115-118 (1990).
7. C. D. Thomas. Nature 347, 237 (1990).
8. I. D. Hodkinson, D. Casson. Biol. J. Linn. Soc. 43, 101-109 (1991).
9. I. D. Hodkinson. J. Trop. Ecol. 8, 505-508 (1992).
10. H. M. AndrŽ , M. I. Noti, P. Lebrun. Biodiver. Conserv. 3, 45-56 (1994).
11. N. E. Stork, K. Gaston. New Sci. 127, 31-35 (1990).
12. J. A. McNeely, K. R. Miller, W. V. Ried, R. A. Mittermeier, T. B. Werner. Conserving the World's Biological Diversity. IUCN, Gland, Switzerland (1990).
13. M. E. SoulŽ . Ann. Mo. Bot. Gar. 77, 4-12 (1990).
14. K. J. Gaston, R. M. May. Nature 356, 281-282 (1992).
15. R. M. May, S. Nee. Nature 378, 447-448 (1995).
16. L. Tangley. BioScience 40, 633-636 (1990).
17. R. G‡mez. Trends Ecol & Evol. 6, 377-378 (1991).

18. D. H. Janzen. Amer. Entomol. 159-171 (1991).
19. D. H. Janzen. In Proceedings of the Norway/UNEP Expert Conference on Biodiversity, Trondheim, Norway (D. T. Sandlund, P. J. Schei, eds), pp. 100-113, NINA (1993).
20. J. T. Longino. Biol. Inter. 28, 3-13 (1994).
21. I. Oliver, A. J. Beattie. Conserv. Biol. 10, 99-109 (1996).
22. N. E. Stork. J. Trop. Ecol. 7, 161-180 (1991).
23. F. C. Thompson. Insect Coll. News 9, 2-4 (1994).
24. D. H. Janzen. Insect Coll. News 7, 24 (1992).
25. I. Oliver, A. J. Beattie. Conserv. Biol. 7, 562-568 (1993).
26. A. J. Beattie, I. Oliver. Trends Ecol & Evol. 9, 488-490 (1994).
27. I. Oliver, A. J. Beattie. Conserv. Biol. 11, 575-576 (1997).
28. T. L. Erwin. In Tropical rainforest ecology and management (S. L. Sutton, T. C. Whitmore, A. C. Chadwick, eds), pp. 59-75. Blackwell Science, Oxford(1983)
29. M. Edwards, D. R. Morse. Trends Ecol & Evol. 10, 153-158 (1995).
30. P. J. D. Weeks, K. J. Gaston. Biodiver. Conserv. 6, 263-274 (1997).

