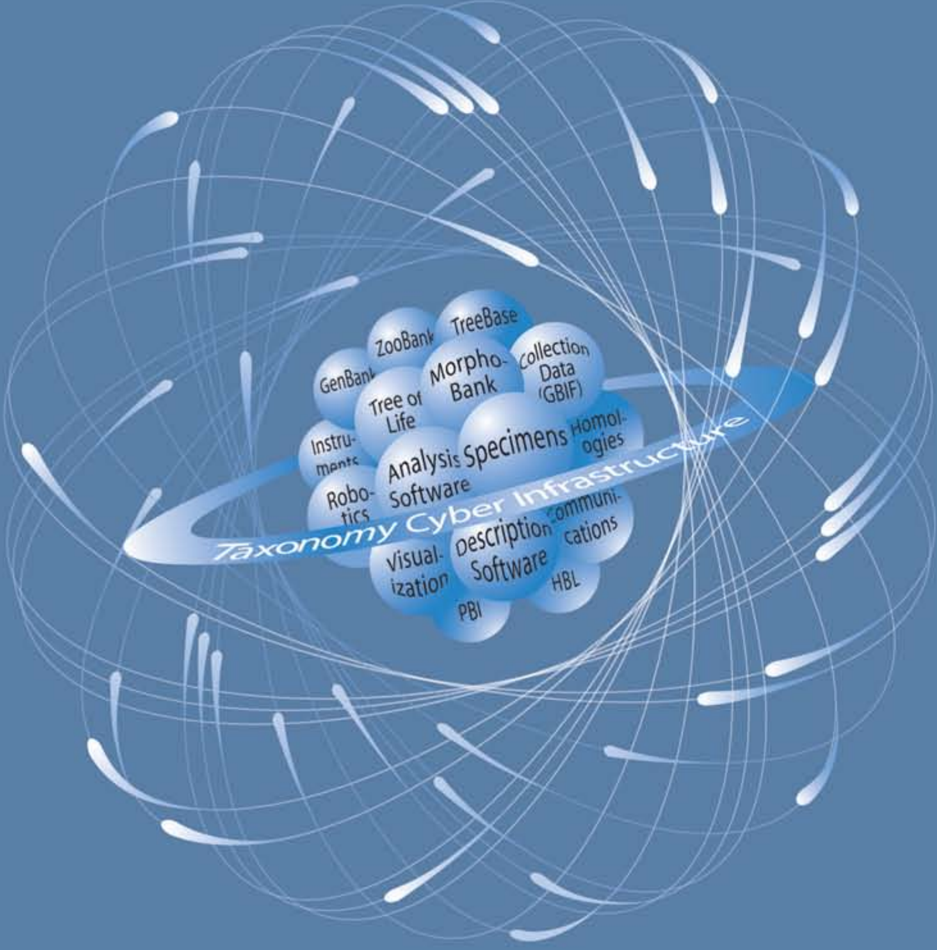
The New Taxonomy



THE Systematics ASSOCIATION

CRC

Edited by
Quentin D. Wheeler

The Systematics Association Special Volume Series 76

The New Taxonomy

The Systematics Association Special Volume Series

Series Editor

Alan Warren
Department of Zoology, The Natural History Museum
Cromwell Road, London SW7 5BD, UK

The Systematics Association promotes all aspects of systematic biology by organizing conferences and workshops on key themes in systematics, publishing books and awarding modest grants in support of systematics research. Membership of the Association is open to internationally based professionals and amateurs with an interest in any branch of biology including palaeobiology. Members are entitled to attend conferences at discounted rates, to apply for grants and to receive the newsletters and mailed information; they also receive a generous discount on the purchase of all volumes produced by the Association.

The first of the Systematics Association's publications *The New Systematics* (1940) was a classic work edited by its then-president Sir Julian Huxley, that set out the problems facing general biologists in deciding which kinds of data would most effectively progress systematics. Since then, more than 70 volumes have been published, often in rapidly expanding areas of science where a modern synthesis is required.

The *modus operandi* of the Association is to encourage leading researchers to organize symposia that result in a multi-authored volume. In 1997 the Association organized the first of its international Biennial Conferences. This and subsequent Biennial Conferences, which are designed to provide for systematists of all kinds, included themed symposia that resulted in further publications. The Association also publishes volumes that are not specifically linked to meetings and encourages new publications in a broad range of systematics topics.

Anyone wishing to learn more about the Systematics Association and its publications should refer to our website at http://www.systass.org

Other Systematics Association publications are listed after the index for this volume.

The Systematics Association Special Volume Series 76

The New Taxonomy

Edited by

Quentin D. Wheeler

Arizona State University

Tempe, U.S.A.

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover art computer graphic by Frances Fawcett, Ithaca, New York.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8493-9088-3 (Hardcover)

Library of Congress Cataloging-in-Publication Data

The new taxonomy / editor, Quentin D. Wheeler.
p. cm. -- (The systematics association special volume series)
Includes bibliographical references and index.
ISBN 978-0-8493-9088-3 (hardback : alk. paper)
1. Biology--Classification. 2. Evolution (Biology) I. Wheeler, Quentin, 1954- II. Title. III. Series.

QH83.N49 2008
578.01'2--dc22 2008004860

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

This book is based on papers presented at the 2005 biennial meeting of the Systematics Association in Cardiff, Wales. Some speakers at the symposium, titled 'The New Taxonomy', did not submit chapters, including Dennis Stevenson of the New York Botanical Garden and James Woolley of Texas A & M University, but nevertheless delivered excellent presentations. Some coauthors who were not present at the meetings contributed to chapters in this book.

The symposium had an unusually high level of positive energy, something commented upon by numerous members of the audience. Given the current state of support for and recent history of taxonomy, many meetings lament the inadequacy of funding for taxonomy education and research, species inventories, collection growth and development, opening access to research resources and creating the digital instruments and elements of cyberinfrastructure desperately needed. Taxonomists and various reports about taxonomy in both the UK and USA have pointed out that most of the limited annual funding goes to support molecular phylogenies rather than integrative data-sets, revisions, monographs or improved classifications and names.

In these years of Linnaean celebrations, 2007 and 2008 (the 300th anniversary of the birth of Carolus Linnaeus and the 250th anniversary of the publication of the 10th edition of *Systema naturae,* respectively), it is appropriate that we reflect on the accomplishments, strengths and promise of taxonomy. The Cardiff symposium and this volume invited participants to imagine what a positive future might look like for taxonomy, assuming that its fortunes change. The results have been truly heartening and inspiring.

It was about 40 years ago that the entomologist Howard Ensign Evans (1969) pointed out how little we knew of our own planet as we ventured into the space age. And it has been more than 20 years since E.O. Wilson (1985) captured our imaginations by comparing how little we know of Earth's species compared to subatomic particles or astronomical bodies. Wilson emphasized that we cannot say how many species are living today even within an order of magnitude. This sobering observation is, sadly, as true today as it was then. In spite of international treaties and great attention to deforestation and rates of extinction, we have made very little progress in organizing or funding a serious scientific response to the biodiversity crisis. Regretfully, attempts to do so are perceived as threats to funds available for molecular phylogenetics and conservation biology; rather, they should be welcomed as the advances in science that must logically precede and underpin any long-term strategies to reconstruct the great tree of life or save as much biological diversity as possible.

While there are many biological and environmental sciences that must be funded for a multitude of reasons, it is a simple fact that no investment in science is more urgent, timely or certain to be repaid in leaps of knowledge and understanding than funding taxonomy, taxonomic collections and a taxonomy-specific cyberinfrastructure. This must be done while remaining cognizant of the nonpareil contributions of

Linnaean classifications and names and of their role, when informed by cladistics, as biology's general reference system (Hennig, 1966; Nelson and Platnick, 1981). The greatest biological 'big science' project ever enjoined–an inventory of all the species of a planet launched by Linnaeus in the middle eighteenth century–and one of the greatest theoretical revolutions of the twentieth century–Hennig's phylogenetic systematics–must be rejoined if we are to learn enough about our world's diversity to understand its evolution and functions and if we are to pass a legacy of specimens and knowledge on to future generations. Hennig's revolution was stopped short of its potential (Nelson, 2004), as was Linnaeus's project. This volume is a tribute to them both and a plea to society to reinvest in the most fundamental of all biological sciences.

The introductory chapter presents the thesis that the decline in prestige and support for taxonomy can be traced from the beginnings of the New Synthesis and, in particular, from Huxley's *The New Systematics,* published as the first special volume of the Systematics Association in 1940. Thus, it is poetic justice that this volume, *The New Taxonomy,* is now the latest special volume of the Systematics Association. This may be seen as a deliberate attempt to reverse the erosive effects of the new systematics and to modernize and spark a revival in descriptive taxonomy before it is too late to explore and document the millions of species that are the results of billions of years of evolution. This book is a call to arms for the taxonomy and museum communities to come together and to organize, plan, innovate and initiate the most ambitious period of exploration in the long history of taxonomy.

I thank the Systematics Association for supporting both the symposium and the book, the contributors who have shared their passion and vision for taxonomy and the audience at the Cardiff meetings for inspiring discussion. And thanks also to my wife, D. Marie Wheeler, for assistance and encouragement while I was editing the volume.

Quentin D. Wheeler
Tempe, Arizona

REFERENCES

Evans, H.E. (1969) *Life on a Little-Known Planet,* Dutton, New York, 318 pp.

Hennig, W. (1966) *Phylogenetic Systematics,* University of Illinois Press, Urbana, IL, 263 pp.

Huxley, J. (1940) *The New Systematics,* Oxford University Press, Oxford, 583 pp.

Nelson, G. (2004) Cladistics: Its arrested development. In *Milestones in Systematics* (eds D.M. Williams and P.L. Forey), CRC Press, Boca Raton, FL, pp. 127–147.

Nelson, G. and Platnick, N.I. (1981) *Systematics and Biogeography: Cladistics and Vicariance,* Columbia University Press, New York, 568 pp.

Wilson, E.O. (1985) The biological diversity crisis: A challenge to science. *Issues in Science and Technology,* 2: 20–29.

The Editor

Quentin D. Wheeler is Virginia M. Ullman professor of natural history and the environment, director of the International Institute for Species Exploration and vice president and dean of the College of Liberal Arts and Sciences at Arizona State University in Tempe. He conducts research on beetles (Insecta: Coleoptera) and has long had an interest in the theory and practice of taxonomy and the role of taxonomy in the exploration and study of biological diversity. As founding director of the International Institute for Species Exploration, he is overseeing a convergence of descriptive taxonomy, phylogenetics, and computer science and engineering to overcome impediments to rapid progress in all phases of species exploration.

Wheeler joined the faculty of Cornell University in Ithaca, New York, following completion of his doctoral degree at The Ohio State University. He was a professor at Cornell in both entomology and plant biology and served as chair of the department of entomology as well as director of the Liberty Hyde Bailey Hortorium. Subsequently, Wheeler served the U.S. National Science Foundation as a program officer and as a division director for 3 years. Wheeler was the first foreign appointed keeper of entomology at London's Natural History Museum. He has authored nearly 100 scientific publications and was coeditor of *Fungus–Insect Relationships: Perspectives in Ecology and Evolution* (with M. Blackwell, Columbia University Press, 1984), *Extinction and Phylogeny* (with M.J. Novacek, Columbia University Press, 1992) and *Species Concepts and Phylogenetic Theory: A Debate* (with R. Meier, Columbia University Press, 2000).

Contributors

James L. Edwards
Global Biodiversity Information Facility
Copenhagen, Denmark

Nico M. Franz
Department of Biology
University of Puerto Rico
Mayaguez, Puerto Rico

Sandra Knapp
Department of Botany
The Natural History Museum
London, England

Norman MacLeod
Department of Palaeontology
The Natural History Museum
London, England

Rudolf Meier
Department of Biological Sciences
National University of Singapore
Singapore

Lawrence M. Page
Natural History Museum
University of Florida
Gainesville, Florida

Robert K. Peet
Department of Biology
University of North Carolina
Chapel Hill, North Carolina

Andrew Polaszek
International Commission on Zoological Nomenclature
The Natural History Museum
London, England

Richard Pyle
Hawaii Biological Survey
Bishop Museum
Honolulu, Hawaii

Malcolm J. Scoble
Department of Entomology
The Natural History Museum
London, England

Larry Speers
Global Biodiversity Information Facility
Copenhagen, Denmark

Alan S. Weakley
The University of North Carolina Herbarium
University of North Carolina
Chapel Hill, North Carolina

Quentin D. Wheeler
International Institute for Species Exploration
Arizona State University
Tempe, Arizona

Doug Yanega
Entomology Research Museum
University of California
Riverside, California

1 Introductory

Toward the New Taxonomy

Quentin D. Wheeler

CONTENTS

INTRODUCTION

When Julian Huxley edited *The New Systematics* in 1940, he may not have appreciated fully the extent to which it would provide the battle cry for those who would dilute, detract from and eventually decimate taxonomy. To the contrary, he began his introduction by stating that, in advocating the 'New Systematics', he nonetheless acknowledged value in the old:

> To hope for the new systematics is to imply no disrespect for the old. It has been largely the rapid progress made by classical taxonomy itself that has necessitated the introduction of new methods of analysis, new approaches to synthesis. The very success of taxonomists collecting material from all parts of an organism's range, in separating and naming groups, and of drawing ever finer distinctions has thrown up a number of general questions which must be answered if taxonomy is to find principles which will enable it to cope with the vast burden of its own data, and to advance to the status of a fully fledged sub-science in which observation and theory, deduction and experiment, all contribute to progress. (p. 1)

Unfortunately, Huxley and his contemporaries conflated pattern with process and supposed, as was the prevailing view of the emerging 'New Synthesis', that the former (e.g. species or clades) could be understood merely as the latter (i.e. evolutionary processes) progressed over sufficient periods of time. He was confusing the methods and goals of the emerging science of population genetics with those of the established science of taxonomy; this is understandable to the extent that questions of species and speciation do intersect and necessarily overlap to some degree along that line between micro- and macro-evolution (Nixon and Wheeler, 1992). The epistemological bases of population genetics–an experimental biology–and of taxonomy–a historical and comparative but non-experimental biology–are simply incompatible. To do either well, one simply cannot use a single approach to both.

This is not to say that one is better or more important than the other. They are simply focussed on very different aspects of biology and necessitate very different assumptions, theories and methods. It is necessary to separate pattern from process in order to do justice to either (Gaffney, 1979; Eldredge and Cracraft, 1980; Nelson and Platnick, 1981). Whether Huxley would have chosen the same course had he known it would undermine the most fundamental subdiscipline of biology or might have advocated a new science of (the processes of) speciation and left taxonomy to its appointed duties is moot; the effect on taxonomy was profound and it helped reinforce a tone in biology that exalted experimentalism over descriptive work and contributed to a disdain and neglect of the latter.

Today, we suffer from the accumulation of 70 years of such confusion and bias. The same oil-and-water mixture of population genetics and taxonomy persists to the present, with 'phylogenetic' analyses done at the population and even individual organism level as though concepts such as 'monophyly' or 'synapomorphy' had logical standing below the species level (see Hennig, 1966; Nelson and Platnick, 1981). Mayr's 1942 *Systematics and the Origin of Species* succeeded in further confounding these incompatible approaches, effectively diminishing the stature of taxonomy and higher level phylogeny study in order to make taxonomy appear to be a modern, genetic-minded (in the parlance of the time, 'population thinking') science (Wheeler, 1995a). There was enormous implied peer pressure associated with Mayr's redefinition of taxonomy: Who would want to be accused of denying modern experimental methods or genetic theory or of pigeonholing specimens blissfully ignorant of advances in evolutionary theory? No such denial of experimentalism or genetics or evolutionary theory is associated with doing taxonomy, of course, but the potential stigma was and remains a frightening one. Just as denying experiments to most subdisciplines of biology would necessarily detract from the credibility of their observations and conclusions, so imposing experimentalism upon a field that is, instead, comparative and concerned with historical *patterns* rather than experimentally measurable *processes* necessarily weakens the quality of its science.

The rigorous nature of post-Hennigian hypothesis testing is truly impressive (e.g. Gaffney, 1979; Nelson and Platnick, 1981) and need not be disguised in the garb of experimentalism to deserve respect. Unfortunately, the great revolution inspired by Hennig was stopped before its completion–before its positive impacts on taxonomy were fully realized (Nelson, 2004; Williams, 2004; Wägele, 2004). The current popularity of neophenetics and an effectively single source (molecular) of evidence are diminishing the reliability and conceptual impacts of taxonomy while the widespread divorce of phylogenies from formal Linnaean classifications and names has derailed progress towards Hennig's general reference system. At the same time a pseudoscience based on quantitative manipulations of questionable phenetic data has focused on cladograms at the exclusion of character analysis. Reversing this trend will require a return to the refinement of Hennig's theories and methods and the integration of diverse available data (Will et al., 2005).

Huxley's *The New Systematics* did a great deal for birthing the new field of population biology and for this reason alone it was an undeniably good thing. It was essential that biologists learn about the origin of mutations, the forces shaping changes in gene frequencies within and among populations and the varied factors related to speciation. It was not essential that this positive advance for population biology, genetics and micro-evolution be realized at the expense of taxonomy and the equally legitimate scientific goals associated with systematic biology (e.g. Cracraft, 2002; Wheeler, 2004; Page et al., 2005). Rather than recognizing the need for a new science associated with mutation, selection and shifting gene frequencies–that is, the study of the *processes* of speciation, Huxley and others tried to redefine the existing science of taxonomy to meet this need. This untenable mongrel science continued, eventually diverging into a very rigorous and well-funded population biology and a bastardized, syncretistic, poorly funded 'evolutionary' taxonomy that suffered from this unfortunate confusion of processes and patterns (Eldredge and Cracraft, 1980).

Huxley had also the seemingly laudable ambition of a unified biology. This was a well-intended goal that could only have been realized by acknowledging the difference between taxonomy and phylogeny on the one hand and experimental subdisciplines of biology on the other and seeking to integrate their findings, rather than their assumptions and methods. Once this theoretically confused mixture was concocted, something had to give. With the avalanche of theoretical and empirical breakthroughs in genetics in the 1930s and 1940s–and with taxonomy decades away from its own essential theoretical advances (i.e. Hennig, 1966; Nelson and Platnick, 1981; Schoch, 1986; Schuh, 2000)–taxonomy's losses in this unstable mixture were inevitable and might have been foreseen. That the pendulum would swing so far to the experimental side that taxonomy would be suppressed for the better part of a century would have been much more difficult to anticipate.

Hennig (1966), himself a product of his times, reflected more process assumptions in his own writings than were necessary to the integrity of his theories (see, for example, Platnick, 1979; Nelson and Platnick, 1981). The core of his theories nonetheless was sufficient to resurrect taxonomy (temporarily, at least) and to set in motion a series of theoretical and methodological advances that would effectively tease apart pattern from process issues, resulting in making taxonomy far more rigorously testable than at any prior time in its long and illustrious history (e.g. Eldredge and Cracraft, 1980; Nelson and Platnick, 1981; Schoch, 1986; Schuh, 2000). The full negative consequences of a recent relapse are yet to be generally acknowledged, much less evaluated or appreciated. Molecular phylogeneticists have reintroduced process assumptions (rates of particular mutations, molecular clocks and other assumptions collectively contributing to maximum likelihood and other models). 'Cladistic' analyses are applied at levels where their theoretical assumptions never suggested they could be trusted (i.e. at the population and even individual specimen levels) to reveal historical patterns with fidelity and phenetic methods that bear a dubious correlation with special similarity (Hennig, 1966) at best. The fundamental phenetic equation suggesting that community of similarity equals community of descent was refuted in the context of morphology (e.g. Wheeler, 1981; Ridley, 1986; Schoch, 1986); a similar presumption in regard to DNA sequence data is yet to be fully examined in either theory or practice, although such methods are widely and uncritically applied.

The extent to which either Huxley or Mayr was concerned by the continued welfare of taxonomy is yet to be evaluated by historians. Regardless of the ultimate answer to such questions, it seems evident to me that, unlike Huxley, Mayr, in his support for the New Systematics, did intend disrespect to the old. Contrast, for example, the following two passages from his 1942 book, *Systematics and the Origin of Species*:

> The old systematics is characterized by the central position of the species. No work, or very little, is done on infraspecific categories (subspecies) … The major problems are those of a cataloguer or bibliographer, rather than those of a biologist. (p. 6)

> The new systematics may be characterized as follows: The importance of the species as such is reduced, since most of the actual work is done with subdivisions of the species, such as subspecies and populations. (p. 7)

Mayr was not unique, but merely among the most vocal, relentless and effective spokesmen for an expanding notion of what is meant by the 'new' systematic biology. Even earlier, Ferris had claimed in 1928 that systematic biology is 'in its broadest implications essentially synonymous with the study of organic evolution'. While one might have imagined that laying claim to all of evolutionary biology would bring untold riches to taxonomy (the charitable view of what Mayr and others had in mind in their new and widened definition), it had quite the opposite effect. Instead, support was focused on 'evolutionary biology' *sensu stricto* (increasingly narrowly seen as experimental genetics and micro-evolution) and taxonomy increasingly marginalized in the process. Darwin's evolution included both the origin of mutations and species and higher classifications that he foresaw could ultimately mirror the genealogical history among species. In the wake of the New Systematics, textbooks would, by the middle of the twentieth century, define evolution narrowly as no more than changes in gene frequencies! What happened to the origin of species, not to mention of higher taxa? The not so subtle message was that fields like taxonomy and paleontology, concerned with larger scale comparisons of patterns among species, were something other than, or at least less than, evolution. Such process bias, combined with an overemphasis of the role of experiments in modern science, undermined the very foundations of taxonomy. The depths of this functional view of biology at the exclusion of the equally credible historical, pattern-based view are seen today in the organization of university departments. The 'ologies' are out of fashion and it is widely accepted that various functional themes are sufficient to reflect the breadth of biology.

Mayr's arguments might not have had so much influence had they not been offered–cleverly, I admit–at the crest of this wave of enthusiasm for experimentalism that had been building since the early eighteenth century in France. The dark genius in Mayr's book–an unkind historian might say of his career–lay in a clever rhetoric that made denial of the New Systematics tantamount to a rejection of modern experimental biology and genetics. It was not by accident that he chose to call his preferred species concept the *biological* species concept rather than, say, the *interbreeding* or the *genetic* species concept, which might have been more precise and appropriate, particularly in contrast to concepts that came before. To choose any concept of species other than his own–one that rested upon modern experimental population genetics (in its purest form, experiments upon interbreeding populations)–was, rhetorically at least, to adopt a concept that was nominally non-biological.

In a similar maneuver, Mayr characterized the New Systematics as being concerned with evolution and the old taxonomy merely with subjectively naming and classifying species. To reinforce the modernity and experimental ties of the New Systematics, he made it clear that the interests of the New Systematics lay at the level of population and, therefore, at a level open to experimentation. To be modern was to accept 'population thinking' as though such an appreciation of population level phenomena were not possible in the context of taxonomy. The 'old', by comparison, focused upon grouping species into higher categories that were products of history and that lay outside the realm of experiments. It was an easy sell in the early 1940s, in the midst of great excitement about advances in experimental branches of biology, including genetics, physiology and others. To fail to endorse the New Systematics was to reject 'population thinking': What could be more retrograde than to step away from the population thinking that had just begun to explain mysteries left unanswered by Darwin?

Another use of words would further contribute to the neglect of the core activities of taxonomy. New Systematists subordinated taxonomy to systematics, relegating taxonomy to the rules and rote practice of classification while systematics was concerned with evolution. The implied message was clear and effective. Taxonomy involved arbitrary bookkeeping and pigeonholing practices and legalistic wrangling over scientific names, while systematics was experimental and intellectually exciting and expansive. Such definitions were endorsed by, among many others, the eminent palaeontologist George Gaylord Simpson (1961): 'Systematics is the scientific study of the kinds and diversity of organisms and of any and all relationships among them' (p. 7) and 'taxonomy is the theoretical study of classification, including its bases, principles, procedures, and rules' (p. 11).

In fairness to Mayr, taxonomy was susceptible to his rhetoric for at least two reasons. The first, not to be underestimated, we have mentioned already. Biologists had long slaved in the shadow of the 'hard' sciences–physics and chemistry–and this new-found (if I may refer to the eighteenth century as such) emphasis on experimentalism was raising their stock among sciences to be taken seriously. Improved statistical measures were rapidly expanding confidence in biological research results. Merely mentioning that taxonomy was non-experimental might have been enough to tear away much of its foundations; doing so explicitly and offering an experimental alternative, however, was particularly effective.

The second was an answer to a challenge posed by Charles Darwin. It had been obvious to readers of *On the Origin of Species* that Darwin and other mid-nineteenth century biologists lacked the necessary theories to explain the distribution of characters and to reconstruct evolutionary history in a way that was testable and not merely anecdotal or, at best, read from a fossil record widely acknowledged to be fragmentary and incomplete. How phylogeny could be reconstructed rigorously was the sister mystery to the (molecular) mechanisms of inheritance. Both geneticists and taxonomists had searched for answers to these dual dilemmas that were the great residual enigmas of Darwin's theory. The geneticists had gotten there first and to the victors go the spoils. Taxonomists could either continue to diligently search for the theory that could make their science (that is, the science of species and higher taxa) rigorous or they could jump ship and ride on the success of population genetics.

Mayr and New Systematists chose the latter–the path of less resistance–while others, including Hennig (1966), carried on:

> If in this struggle for survival biological systematics has recently lost ground to other and, as is often heard, younger and more modern disciplines, this is not so much because of the limited practical or theoretical importance of systematics as because systematists have not correctly understood how to present its importance in the general field of biology, and to establish a unified system of instruction in its problems, tasks, and methods. (p. 1)

Hennig succeeded to a very great extent, to which witness the ubiquitous reliance on phylogenies today in virtually every branch of biology from molecular genetics to ecosystem science. His criticism of taxonomists' inept advocacy for their own science's practical and theoretical strengths rings true still, not for phylogeny in and of itself, but rather for what non-taxonomists refer to as 'descriptive' aspects of the field. Hennig's general reference system was not intended to be merely a branching diagram. That is fine, of course, if you are a biologist concerned with some static number of known species, but wholly inappropriate to the broad exploration of biological diversity. His reference system was not a diagram but rather a phylogenetic classification and a set of Linnaean names that correspond to various levels in that classification. It was also an explicit and precise and testable account of the complex characters that ultimately make such higher patterns of such great interest to science. It is time to complete the Hennigian revolution and for taxonomists to rise to his challenge to become effective proponents for the whole of taxonomy, not merely phylogenetic analysis (Wheeler, 2004).

Anyone who doubts the extent to which the rest of taxonomy has become marginalized need only read two recent papers by Felsenstein:

> The focus of systematics has shifted massively away from classification: it is the phylogenies that are central, and it is nearly irrelevant how they are then used in taxonomy. (2001, p. 467)

> The delimitation of higher taxa is no longer a major task of systematics, as the availability of estimates of the phylogeny removes the need to use these classifications. Thus the outcome of the wars over classification matters less and less. (2004, p. 145)

Felsenstein's lack of appreciation for taxonomy and the importance of classifications as a general frame of reference for biology and of nomenclature as the foundation of a precise and stable language for biological diversity is staggering. Felsenstein is, however, exemplary of the branch of population genetics that adopted the terminology and quasi-analytical aspects of phylogenetic systematics while apparently never grasping the theoretical importance of distinguishing tokogeny from phylogeny (Hennig, 1966) or of precise scientific terms clearly linked to concepts embedded in a formal Linnaean classification. For taxonomists concerned with the exploration and characterization of all life on Earth and not merely the connections among populations of a few species or relative relationships expressed graphically in a cladogram, the narrowness of Felsenstein's understanding of taxonomy is breathtaking.

Perhaps some of the damage to taxonomy could have been mitigated had Hennig's ideas received wide attention sooner. Hennig had worked on his theories at the same time that the New Systematics was diverting attention away from the core questions of systematics. They did not have much international impact, however, until a manuscript was translated and published in English in 1966. By then, a great deal of damage had been done. All the same, Hennig's *Phylogenetic Systematics* had solved Darwin's last great mystery and set forth a comprehensive, rigorous, theoretical and methodological framework for reconstructing phylogeny and critically testing that historical pattern in detail. The tree-like diagrams that had been introduced by Häckel (see Figure 1 in that work) that were referenced in later editions of the *Origin* were no longer merely graphic summaries of speculations of learned naturalists. They were visual presentations of testable hypotheses that make explicit predictions about the distributions of characters among species and monophyletic groups of species. Thanks to Hennig and his followers, cladograms could be critically tested and then translated directly into Linnaean classifications and names. Taxonomy, with its focus on species and groups of species, was, finally, as scientific as the population thinking of the New Systematics or the experimental fields of biology. The importance of a reliable reference system cannot be overstated. In the words of Eugene Gaffney (1979):

> [I]t seems to me that the most important biological reference system is a phylogeny and that a classification that mirrors a phylogenetic hypotheses (however transitory) is the most useful for systematists and non-systematists. (p. 79)

Hennig had a predictable, positive impact on taxonomy and there was an impressive, if regrettably short-lived, revival in taxonomy for taxonomy's sake. With newfound respect, Linnaean names and classifications not only were suddenly a demonstrably *scientific* form of taxonomy, but also were transformed, as Hennig (1966) had envisioned, into *the* general reference system for biology.

Already severely weakened by the New Systematics, taxonomy could not weather the next wave of fashion in biology: the rise of molecular genetics. As Hennig emphasized, *how* data are analysed is far more important to the rigor of taxonomy than the source of those data. Therefore, the arrival of a new class of data–molecular sequence data–should have been nothing but good news for taxonomy and might have been so, had the lingering shadow of the New Systematics not persisted. Instead, seemingly unconnected developments combined into a perfect storm to once again undermine the stature of and support for taxonomy as an independent science, including global parsimony analysis, availability of DNA sequence data, advances in computational cladistics and a flood of funding to molecular geneticists who had either never read or did not appreciate Hennig's theories.

If my armchair analysis of events has not offended serious historians yet, here is where I really climb out on a limb. I suggest that it was, specifically, a confluence of at least five factors that served to undermine the otherwise entirely positive impact of Hennig. Were it not for these events, the trajectory of Hennig's phylogenetic systematics would have been (I suggest) the widespread adoption not only of phylogenies in biology but also of formal classifications that were both phylogenetic and

Linnaean. In this alternative world, taxonomy would have been widely respected and well funded as biologists in general understood that, in order to store and retrieve observations, communicate information about species or interpret in evolutionary context any comparative data across species boundaries, they would turn to phylogenetic Linnaean classifications and nomenclature. To recap what has been said thus far and tie these historical threads together, let us examine these factors that derailed the remarkable inertia of Hennig's theories.

First, there was the lingering suspicion, held by most experimental biologists, that taxonomy was somehow not really science; after all, it was explicitly non-experimental and generations of biologists had been educated at a time when much of taxonomy and many of its ideas were not visibly testable. Second, there was a shift in preference from individual character analysis to 'global' character analysis. That is, *a priori* determination of which characters were apomorphic–a necessary step to discern which characters met the requirements of Hennig's 'special similarity'–could be avoided by analysing all characters simultaneously. They were then assessed in a minimum distance network and subsequently rooted using a designated out-group. The removal of an emphasis on character analysis denied to taxonomy its greatest and most visible intellectual content. Trees were needed by other biologists to interpret their own observations in a phylogenetic context. Devoid of a character focus, taxonomy began to look like and behave more as a service than an independent science. Third, a disproportionate focus was put on constructing cladograms from data matrixes and on the selection of one cladogram from among a set of equally parsimonious ones–as opposed to the previous emphasis on the careful analysis of individual characters and explicit polarity hypotheses. Fourth, improved computer algorithms could cope with ever larger data-sets. Fifth, relatively inexpensive and increasingly abundant molecular sequence data arrived.

Once the challenge was perceived to be finding the shortest tree from a 'given' data matrix or the preferred tree among equally parsimonious ones and the emphasis was removed from individual characters, then populating character matrixes became increasingly easy as DNA sequences flowed freely from multiplying numbers of laboratories. Relieved of the need to conjecture about polarity, focusing on individual characters was less obviously of paramount concern. As soon as DNA sequence data became available, the number of molecular 'characters' quickly swamped those from morphology in matrixes, further diminishing the contributions of morphology. Rather soon it became obvious, at least to 'phylogenetic biologists', that the need for morphology was diminishingly small. Cladograms could be generated faster and more easily by simply relying on sequence data, so they were. Ironically, while cladograms could be produced rapidly and in profusion, without detailed analysis of characters, we quickly lost sight of why we wanted the cladograms in the first place: to interpret and understand the evolutionary patterns that explain the origin and diversification of complex characters.

Just as users of taxonomic names simply want to identify species and know what to call them and do not appreciate the science behind species testing that is taxonomy, so many users of cladograms simply want to know phylogenetic relationships so that they can interpret some phenomenon or data that they are studying. They do not share the interest in character transformations, homology, synapomorphy or of cladogensis in and of themselves or the historical biogeographical patterns that can be studied as a result. Contemporary molecular phylogenetics is much more like the applied service aspect of alpha taxonomy than the research paradigm advocated by Linnaeus and Hennig and thousands of taxonomists. Just as identification services were and remain a noble, necessary and worthy contribution to science and society, so is the provision of phylogenetic patterns. However, in each case the most profoundly interesting aspects of taxonomy and the most inspired and best comparative, monographic and macro-evolutionary work come from taxonomy for its own sake. However, taxonomy is grossly misunderstood by a modern biology that is so steeped in an ecological perspective and experimentation that the very existence of two different hierarchies seems to have been forgotten (Wheeler, 1995b).

Most biologists think of functional hierarchies progressing from individuals to demes to ecosystems, even though these are not hierarchical in a strict mathematical sense (e.g. Woodger, 1937). The other and true hierarchy is historical, rather than ever more encompassing groupings of functional assemblages of organisms. That hierarchy is phylogeny, the relative relationships among species due to a shared history of descent with modification and speciation.

Molecular systematics enjoyed the same advantage of success by association that the New Systematics had used to its advantage decades earlier. Molecular techniques were new, expensive and technologically impressive. To paraphrase a comment made at a Washington, D.C., meeting of the American Association for the Advancement of Science (AAAS) by Professor Edward O. Wilson of Harvard University, molecular labs are not flooded with money because their data are better; molecular data are perceived to be better because they receive so much funding. That funding reflects, too, an association by technique with biomedicine, whose obscene levels of funding cast a long shadow on basic biology. The House of Lords report in 2002 noted that most funding designated for systematics in the UK in fact had supported, for the previous decade, studies that resulted in cladograms but almost never in improved classifications or scientific names. The same was true in the USA. There are some fascinating parallels between molecular systematics and the New Systematics that should send chills down the spine of anyone who appreciates the unique and vital role played in biology by the science of taxonomy.

Advocates of the New Systematics called themselves systematists to distance themselves from (old) taxonomy (by 1970 this had morphed further to 'biosystematics', presumably to increase the distance even farther); they justified their superiority by emphasizing their modern 'population thinking'; they gained most of their funding by implying that they were the modern alternative to taxonomy, successfully redirecting funds from taxonomic work; and they propelled their movement forward by tightly associating it with modern genetics so that to attack one was implicitly to attack the other. No one wants to be politically incorrect now any more than one did in 1942. Consider now the parallels with contemporary workers who frequently call themselves either 'molecular systematists' or 'phylogenetic biologists' to distance themselves from phylogenetic systematists of the Hennig era. (O'Hara, 2002, suggests that phylogenetic biology, like population biology, is a new science arising largely out of systematics but with different goals.) They justify their superiority by emphasizing 'tree thinking', gain most of their funding by redirecting monies allocated to systematics into projects that reconstruct phylogeny without improving classifications or names and propel their science forward by association with molecular genetics and biomedical molecular research. Sound familiar?

TIME FOR A CHANGE

With the recognition from other biologists of the need for phylogenetic context, many are complicit with Felsenstein in funding phylogenies that are never translated into formal classifications and Latin names–the ultimate expression and benefit of Hennig's general reference system.

Phylogenetic biology (or molecular systematics or whatever you wish to call it) has begun to wear thin, however. There are reasons to suspect that all is not so objective and easy as the molecular neopheneticists claim. Even if that were not the case, however, there are other concerns. Its putatively self-evident importance was derived from expenditure of the intellectual capital banked by morphological taxonomists for 250 years. It is rapidly becoming a victim of its own success. As it reanalyses data-set after data-set, it is depleting this capital but it is not reinvesting in the study of the complex characters that can be explained by its results or in the basic exploration and enumeration of Earth's species that provide the raw material of cladograms. A cladogram is not interesting in and of itself. If I tell you that a cladogram depicts relationships among three undescribed species about which I know nothing except the order of amino acids in a few segments of DNA molecules, will you care that species A is more closely related to species B than either is to species C? The cladogram is only interesting because and to the extent that it allows us to understand and interpret in a historical context the origin and diversification of complex characters whether they are morphological, behavioural or whatever. Keep in mind that we estimate that 90 per cent of the species on Earth are unknown to science; we know absolutely nothing about their morphology, behaviour, ecology or geography. Molecular phylogenies for 90 per cent of the species on Earth are largely pointless, except where they group new species with known species for which, once again, taxonomists and morphologists have already documented characters of interest.

There is a growing recognition that we live in the early days of a major biodiversity crisis and that an enormous number of species and clades will soon become extinct (Wilson, 1985, 1992; Raven, 1995). It is vitally important that we explore, describe, name and classify these species as the last possible evidence for a great deal of evolutionary history. The value of gathering these species and observations about their morphology, ecology, geography and behaviour is inestimable, yet support for the most basic 'descriptive' taxonomy continues to be drained almost entirely into phylogenetic biology. Clearly, funders of biological research and self-described phylogenetic biologists have lost the plot. Phylogenetic theory was important because it transformed already useful Linnaean classifications and names into a predictive and optimally efficient information storage system, language and conceptual *general reference system* for biology. As Hennig reasoned, evolutionary history is the only thread woven through the whole of life on Earth; even the processes of evolution, such as various kinds of selection forces, are unevenly shared. It is only the *patterns* of character distributions–the stuff of taxonomy–that underpin a unitary view of life.

That we can construct cladograms faster with DNA data is irrelevant unless there are facts known about things worth explaining!

What is ultimately most interesting about evolution is complexity: the complexity of an orchid flower found attractive by insect pollinators, the complexity of an ecosystem comprising hundreds of thousands of species, the complexity of processes by which natural selection proceeds and the seemingly endless complexity of morphology and its interface with the environment. The Human Genome Project revealed that the human genome comprises only about 30,000 rather than the anticipated 100,000 genes. This suggests that, even at the DNA level, genetic coding is more complex than what sequences alone can tell us. Epigenetic factors are being discovered in association with diseases that should give heart to any die-hard fans of Lamarck, and understanding human genetics and genetic-related diseases will require biological investigations above the level of sequences.

Several authors in *Milestones in Systematics,* edited by Forey and Williams (2004), advocated a returned emphasis on morphology and on explicit character analysis. Thinking about individual complex characters has merit in itself–before a cladistic analysis and particularly afterward when results seem surprising. Again, Hennig (1966) was prophetic in his rule of thumb of questioning cladistic results that were at odds with what a consideration of complex characters had first suggested. It is imperative for the intellectual content of taxonomy and of evolutionary biology that this sage advice is heeded. The name of the evolution game is complexity and it is written in complex characters that deserve great thought and interpretation as objects of intense study. It seems likely that, as molecular systematics matures, it will become obvious that the most interesting ‘characters’ are not in fact sequences of base pairs but instead emergent complex molecular characters, whether that refers to molecule folding patterns, genes, proteins or sets of interacting genes. We seem stuck in the usual early phase of a new technology when time is spent simply marvelling over new parlour tricks; it is time that molecular evidence be regarded for what it is: data. It is only data and as such it is no better than data from comparative morphology, ontogenetics, ethology or palaeontology.

As an aside, molecular data are possibly *less* than data from these other sources; it is, again, complexity that adds layers of historical information to morphological characters that can be studied at increasingly fine levels of structure, at the level of ontogeny and, ultimately, at the molecular developmental level. The sooner we return the logical focus to what matters most, the better: the loss of species and clades; the need to explore and chart the species of the biosphere; the need to study and document as many characters as possible, especially those that are complex; the need to integrate diverse data in taxonomy; and the need to return support to taxonomy, including comparative morphology and formal classification and nomenclature.

In everyday use ‘taxonomy’ and ‘systematics’ used to be applied more or less synonymously. Then Mayr and others succeeded in defining systematics in a framework of evolutionary theory, relegating taxonomy to little more than biological bookkeeping. More recently, ‘systematics’ has come to be used essentially for phylogenetics in the sense of cladistic analyses divorced from classification and nomenclature. In my view the ultimate aim of both activities, however they are defined, is to provide the most useful classification system–that is, Hennig’s general reference system. Mayr's inversion of priorities, elevating 'systematics' to mean experimental evolutionary biology and 'taxonomy' the technical application of rules of classification and nomenclature, made the former seem new, sexy and relevant and the latter pedantic and stultified. Like others who have seen taxonomy as the broader field (e.g. Mason, Crowson and others; see Wheeler, 1995a), I see the end goal as a classification and its associated names. The subdiscipline called systematics is concerned with reconstructing phylogeny so that it can inform such classifications and their associated formal names. This relationship perhaps will help us keep our eye on the prize: Linnaeus's account of Earth's species, Darwin's big picture of phylogeny and Hennig's general reference system. We should use all relevant data and tools to further our knowledge of Earth's species and their phylogeny and, ultimately, to make Linnaean names as informative as possible and classifications as information rich and predictive as possible.

With gathering clouds of extinction on the horizon, the time has come to set priorities that include putting taxonomy front and centre in our response to the biodiversity crisis. There will be centuries to explore the inner workings of cells and the coding of proteins. There will be no other planets or multibillion-year evolutionary experiments to explore. This is a one-time offer we cannot afford to pass up and the sell-by date for learning our planet's species has expired. As ecologist David Orr (2003) has suggested, let us sort 'should do' from 'can do' and contribute what we alone can to humanity's ultimate understanding of evolution and complex ecosystems: an encyclopaedic knowledge of the origin, diversification, distribution and associations of Earth's species. This requires an effective, focussed and well-resourced taxonomy that respects the full range of possible data, tools and contributions.

The Internet and a multitude of rapidly developing digital technologies have delivered to us a powerful message: The opportunity exists to reinvent taxonomy for the twenty-first century as a larger, more powerful, more efficient, more productive and more integrated science capable of responding to the biodiversity crisis in the most important way possible, by expanding our knowledge and understanding of biodiversity through taxonomic investigations and mapping of the biosphere.

A blue-ribbon panel informed the US National Science Foundation that we are ready for the greatest revolution to date in the information sciences–ready for a powerful new cyberinfrastructure capable of transforming how science is done, communicated and used (Atkins et al., 2003). No science stands to benefit more from this new cyberinfrastructure than taxonomy, which seeks to deal with billions of facts about millions of species; no science is capable of delivering more relevant information to a rapidly changing world than taxonomy (Wheeler et al., 2004). Put simply, it is impossible to understand, detect, manage, monitor, conserve, sustain or even exploit that which is not known to exist, which cannot be identified when we see it and for which we do not have precise scientific words to communicate what we know.

Every phase of taxonomy can benefit from and be accelerated by the development, adoption and application of new digital tools (Wheeler, 2007), from field inventories to the study of specimens and characters, analysis of cladistic relationships, erection of formal classifications, collection curation and species identifications. One effort to prepare the taxonomy and museum communities for this new cyberworld is the Legacy Infrastructure Network for Natural Environments (LINNE), discussed by Lawrence Page in this volume and also in Page et al. (2005). The LINNE vision is of a vast, distributed taxonomy research platform that makes everything needed by taxonomists available at their fingertips, vastly speeding the work of taxonomy and turning the currently fragmented natural history collections of the world into a united research resource.

A glimpse into the taxonomy of the future can be found in the chapters that follow.

AN ILLUSION OF TAXONOMY

Taxonomy in its useful, modern, integrated form is being undermined by proposals that appear naively ignorant of its theories, epistemology, history and traditions. While the discipline expends energy keeping the forces of ignorance at bay, the resources and efforts needed to complete a study of the macro-evolutionary patterns of life on Earth are denied to the generation that needs them most (Wheeler, 2004). Revitalizing taxonomy is the greatest scientific challenge of our time. Knowledge of our world's species can help us and all future generations expand our understanding of the living world and solve environmental and human welfare challenges. Revitalizing taxonomy is the noblest contribution that our generation can make to humankind. No future generation will ever have access to the number and diversity of species that we have. For comparatively modest costs we can provide a legacy of specimens, data, information and knowledge that will inspire and inform all humans that follow us. This will consist of lots of kinds of data, but first and foremost of natural history collections that provide the most fundamental and important record of life on Earth.

This is the greatest biological 'big science' project ever conceived and will require dedication and focus to succeed. We must avoid the mistakes of the past and the diversions of competing agendas, including the get-rich-quick schemes that seek to sell out the goals of taxonomy to once again appear to be modern and pass off taxonomy as something it is not. Among the current diversions that threaten the taxonomic agenda are DNA barcoding, the PhyloCode, and Felsenstein's 'I don't care' school of classification; each has been soundly refuted and no further energy or time should be wasted on such ill conceived ideas, here or elsewhere. For entrée into the literature refuting each, see Franz (2005) regarding Felsenstein; Carpenter (2003), Keller et al. (2003) and Nixon et al. (2003) regarding Phylocode; and Prendini (2005), Wheeler (2005), Will and Rubinoff (2005) and Will et al. (2005) regarding barcoding.

Because a handful of species currently used as model organisms, that are regarded as vectors or pests or that are the subject of biological experimentation can be identified, we labor under an illusion that existing taxonomic information is sufficient. It is not. In nearly every case, we do not yet know the best model organism or the best source for a natural product or process. We subconsciously or consciously constrain our experiments to what can be done rather than insisting that enough taxonomy be completed so that any organism, system or phenomenon can be studied equally well.

That we have a large number of specimens (currently estimated at as many as three billion worldwide) in natural history museums and herbaria is taken as tacit evidence that much of the work of discovering and documenting species is done. Again, this is merely an illusion. These specimens represent a mere fraction of living species and an infinitesimally small sample of the morphological variation and geographical distributions of these species. Taking insects as an example of a large hyperdiverse group, Grimaldi and Engel (2005) estimate that, at most, 25 per cent of living species are known to science. Thus, what we know of such groups is also illusory.

Historically, taxonomists have used an ingenious high through-put approach to testing 'known' species: monography. Monographs (and taxonomic revisions) are comprehensive, periodic studies of all related species in some larger taxon such as a genus or family. The model is a simple, effective one. All species described since 1758 are reviewed in detail based on an examination of all previously known species, all type specimens, as many specimens as available or practicable including unidentified specimens that have accumulated in collections since the last revision or monograph. Testing a single species hypothesis requires comparisons of all available specimens of that species as well as all those closely related to it. These species tests consist of observations that critically test the predicted unique combinations of characters that make each species unique (Wheeler and Platnick, 2000). By doing a large number of species at once, it is possible to avoid much of the work associated with an effective test of one species in isolation. Unfortunately, this kind of work has fallen out of fashion and is simply not being done on a scale appropriate to testing existing species hypotheses, much less confronting the challenge of discovering new species. The result? Specimens in museums are labelled with names that refer to uncorroborated hypotheses that, with every passing year of neglect, less accurately reflect the actual patterns in nature. Slowly, what we 'know' is increasingly less true; collections, data, publications, names and classifications are eroding, quietly leading inexorably toward a disaster.

Will we do what is right and what ought to be done in the midst of a biodiversity crisis and return support to taxonomy? Or will we continue to neglect taxonomy and ignore it until there is a taxonomic train wreck? Unless we assure that we have access to reliable taxonomic information, such disasters are a question of 'when' rather than 'if'. Will it be an incident of bioterrorism? An introduced pest species that devastates a crop species? An invasive species that drives valuable native plants or animals to extinction? A misidentification that renders important experiments useless? Conservation priorities or measures that cannot succeed? Regardless of the details, each such disaster can easily be avoided by investing in taxonomic research.

CONCLUSIONS

It is clear that, for most of the past century, taxonomy has suffered at the hands of competing trends and fashions in biology and that, each time, it has been marginalized, under-resourced and neglected. It is important to recognize this history and the deeply engrained prejudices it reveals if we are to avoid repeating such mistakes again. It is important, too, to recognize that taxonomists, or at least individuals purporting to speak for taxonomists, have been complicit with this neglect. Unless taxonomists and taxonomy institutions (primarily those with significant collections) provide the necessary leadership, taxonomy will remain weak and vulnerable to such trends and expedient alternatives. How can taxonomy be made strong? That, I suggest, is the message of this volume. Virtually every aspect of taxonomy–from field inventories to electronic publications–can be transformed into a larger, more efficient, stronger science than it was before. The collective effect of these new approaches, tools and infrastructures will be to strengthen the science and give us the opportunity to complete the incredibly important work ahead.

No science can contribute more to our knowledge of evolutionary history or biodiversity or to human or environmental welfare than taxonomy; no science stands to lose more from the biodiversity crisis than taxonomy. Thanks to Hennig and other phylogeneticists, we have the theoretical framework to do taxonomy better than ever before and to create the general reference system he envisioned. Thanks to information technologies and molecular genetics, we have powerful new tools to do our work faster and better than before. And thanks to the vision of authors in this volume, we see creative, positive options for the future that will enable taxonomy to reach its enormous unrealized potential.

The time has come to complete the revolution in taxonomy begun by Hennig (1966). Half the battle has been won. Hennig's theories and methods for reconstructing phylogeny are now widespread and integral to nearly every aspect of biology even remotely comparative or evolutionary in perspective. Against these spectacular successes, it should be possible to renew the revolution where it left off and to focus specifically on so-called descriptive aspects of taxonomy: character analysis, formal classifications and names. The result should be an independent science of taxonomy pursuing its unique questions and challenges and a vastly improved source of reliable taxonomic and phylogenetic information for a rapidly growing user community.

Even those forces that have diverted us from the core questions of taxonomy (Cracraft, 2002; Page et al., 2005) can be turned to our advantage. We can envision now a bright future for taxonomy based on advances in data sources, the information sciences and taxonomic theory. The ideas proposed in this volume offer much encouragement and many positive ways forward. We have a chance to reverse the negative recent history of the most fundamental of the life sciences, to arm ourselves with knowledge with which to face an uncertain environmental future and to give the greatest gift to future generations: a documentation of life on the most biologically diverse planet known.

The many benefits of a general reference system described by Hennig are the reasons we must pursue this revitalization of *phylogenetic classification*, the biodiversity crisis is the reason that we must make the *exploration of Earth's species* an urgent and top priority in science and the exploding need for reliable information about the world's species, now deliverable efficiently via the Internet, is the reason we must meld information science and engineering with all aspects of taxonomy.

REFERENCES

Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Mesina, P., Ostriker, J.P. and Wright, M.H. (2003) *Revolutionizing Science and Engineering through Cyberinfratructure. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure,* US National Science Foundation, Arlington, VA, 84 pp.

Carpenter, J. (2003) A critique of pure folly. *Botanical Review,* 69: 79–92.

Cracraft, J. (2002) The seven great questions of systematic biology: An essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanical Garden,* 89: 127–144.

Eldredge, N. and Cracraft, J. (1980) *Phylogenetic Patterns and the Evolutionary Process,* Columbia University Press, New York, 349 pp.

Felsenstein, J. (2001) The troubled growth of statistical phylogenetics. *Systematic Biology,* 50: 465–467.

Felsenstein, J. (2004) A digression on history and philosophy. In *Inferring Phylogenies* (ed J. Felsenstein), Sinauer Associates, Sunderland, MA, pp. 123–146.

Ferris, G.F. (1928) *The Principles of Systematic Entomology,* Stanford University Press, Standford, CA.

Franz, N. (2005) On the lack of good scientific reasons for the growing phylogeny/classification gap. *Cladistics,* 21: 495–500.

Gaffney, E.S. (1979) An introduction to the logic of phylogeny reconstruction. In *Phylogenetic Analysis and Paleontology* (eds J. Cracraftand N. Eldredge), Columbia University Press, New York, pp. 79–111.

Grimaldi, D. and Engel, M. (2005) *The Evolution of the Insects,* Cambridge University Press, Cambridge, 755 pp.

Haeckel, E. (1866) *Generelle Morphologie der Organismen.* G. Reimer, Berlin, 2 vols.

Hennig, W. (1996) *Phylogenetic Systematics*, Illinois University Press, Urbana, 263 pp.

House of Lords (2002) What on Earth? The threat to the science underpinning conservation. Select Committee on Science and Technology. 3rd report. HL Paper 118(i). HMSO, London.

Huxley, J. (1940) *The New Systematics,* Oxford University Press, Oxford, 583 pp.

Keller, R.A., Boyd, R.N. and Wheeler, Q.D. (2003) The illogical basis of phylogenetic nomenclature. *Botanical Review,* 69: 83–110.

Mayr, E. (1942) *Systematics and the Origin of Species,* Columbia University Press, New York, 367 pp.

Nelson, G. (2004) Cladistics: Its arrested development. In *Milestones in Systematics* (eds D.M. Williams and P.L. Forey), CRC Press, Boca Raton, FL, pp. 127–147.

Nelson, G. and Platnick, W. (1981) *Systematics and Biogeography: Cladistics and Vicariance*, Columbia University Press, New York, 567 pp.

Nixon, K.C., Carpenter, J.M. and Stevenson, S.W. (2003) The phylocode is fatally flawed, and the 'Linnaean' system can easily be fixed. *Botanical Review,* 69: 111–120.

Nixon, K.C., and Wheeler, Q.D. (1992) Extinction and the origin of species In *Extinction and Phylogeny* (eds M.J. Novacek and Q.D. Wheeler) Columbia University Press, New York, pp. 119–143.

O'Hara, R.J. (2002) Population thinking and tree thinking in systematics. *Zoological Scripta,* 26: 323–329.

Orr, D.W. (2003) *The Nature of Design,* Oxford University Press, New York, 237 pp.

Page, L.M., Bart, H.L., Jr., Beaman, R., Bohs, L., Deck, L.T., Funk, V.A., Lipscomb, D., Mares, M.A., Prather, L.A., Stevenson, J., Wheeler, Q.D., Woolley, J.B. and Stevenson, D.W. (2005) *LINNE: Legacy Infrastructure Network for Natural Environments*, Illinois Natural History Survey, Champaign, Illinois, 16 pp.

Prendini, L. (2005) Comments on 'Identifying spiders through DNA barcodes'. *Canadian Journal of Zoology,* 83: 498–504.

Raven, P.H. (2004) Taxonomy: Where are we now? *Philosophical Transactions of the Royal Society of London, Series B,* 359: 729–730.

Ridley, M. (1986) *Evolution and Classification: The Reformation of Cladism.* Longman Press, London, 201 pp.

Schoch, R.M. (1986) Phylogeny Reconstruction in Paleontology. Von Nostrand Reinhold, New York, 353 pp.

Schuh, R.T. (2000) *Biological Systematics*, Cornell University Press, New York, 236 pp.

Simpson, G.G. (1961) *Principles of Animal Taxonomy.* Columbia University Press, New York, 272 pp.

Wagele, J.W. (2004) Hennig's phylogenetic systematics brought up to date. In *Milestones in Systematics* (eds D.M. Williams and P.L. Forey), CRC Press, Boca Raton, FL, pp. 101–125.

Wheeler, Q.D. (1981) The ins and outs of character analysis: A response to Crisci and Stuessy. *Systematic Botany*, 6: 297–306.

Wheeler, Q.D. (1995a) The 'Old Systematics': Classification and phylogeny. In *Biology, Phylogeny, and Classification of Coleoptera. Papers Celebrating the 80th Birthday of Roy A. Crowson* (eds J. Pakaluk and S.A. Slipinski), Muzeum i Insytut Zoologii PAN, Warszawa, pp. 31–62.

Wheeler, Q.D. (1995b) Systematics and biodiversity. *BioScience,* supplement, 45: 21–28.

Wheeler, Q.D. (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society of London, Series B,* 359: 571–583.

Wheeler, Q.D. (2005) Losing the plot: DNA barcodes and taxonomy. *Cladistics,* 21: 405–407.

Wheeler, Q.D. (2007) Digital innovation and taxonomy's finest hour. In *Automated Taxon Identification in Systematics: Theory, Approaches and Applications* (ed N. MacLeod), CRC Press, Boca Raton, FL, pp. 9–23.

Wheeler, Q.D. and Platnick, P.I. (2000) The phylogenetic species concept. In *Species Concepts and Phylogenetic Theory: A Debate* (eds Q.D. Wheeler and R. Meier), Columbia University Press, New York, pp. 55–69.

Wheeler, Q.D., Raven, P.H. and Wilson, E.O. (2004) Taxonomy: Impediment or expedient? *Science,* 303: 285.

Will, K.P. and Rubinoff, D. (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, 20: 47–55.

Will, K.P., Mishler, B.D. and Wheeler, Q.D. (2005) The perils of DNA bar-coding and the need for integrative taxonomy. *Systematic Biology*, 54: 844–851.

Williams, D.M. (2004) Homologues and homology, phenetics and cladistics: 150 years of progress. In *Milestones in Systematics* (eds D.M. Williams and P.L. Forey), CRC Press, Boca Raton, FL, pp. 191–224.

Wilson, E.O. (1985) The biological diversity crisis: A challenge to science. *Issues in Science and Technology,* 2: 20–29.

Wilson, E.O. (1992) *The Diversity of Life,* Norton, New York, 424 pp.

Woodger, J.H. (1937) *The Axiomatic Method in Biology,* Cambridge University Press, Cambridge, 174 pp.

2 Networks and Their Role in e-Taxonomy

Malcolm J. Scoble

CONTENTS

INTRODUCTION

Revisionary taxonomists are more often apologists than promoters of their discipline. Yet apologies are unnecessary since describing the diversity of life is arguably the most socially important role of taxonomy. Describing life combines a major social role with good scholarship. Taxonomists need not agonize over whether their research is question-driven, ‘real’ science, or, perhaps, even if all parts should even be demarcated as science. Rather they should ensure that it is high-quality, comprehensive research. Is all archaeology or anthropology science? Is historical research compromised because it is not science? Taxonomy should be celebrated for what it is rapidly becoming—a discipline that is as much about informatics as it is about biology. It may be unfashionable for a field to be descriptive, archive dependent (literature and collections), and concerned with historical data (e.g. species and phylogenies) rather than universal statements. But in the light of taxonomy’s profound engagement with information science, its evident value to a society rapidly depleting its biodiversity and the frequency with which the subject is debated in the scientific press mean there is no reason to treat the field as anything other than leading edge. Taxonomists have progressed by adopting an e-dimension in their work, but deeper and quicker engagement is desirable. To encourage such a transformation requires more than e-networking. It also needs taxonomists to plan their research agendas collectively and then to work in harmony. There are encouraging signs that this change is starting to occur.

There has been no shortage of views about taxonomy since the publication of *The New Systematics* (Huxley, 1940). Most of the discussion has been about the need for taxonomy to change by engaging with the new biology, its concepts and techniques, and to be more question oriented. It is argued frequently that taxonomy is declining in terms of funding and recruitment. I take the view that while taxonomy should indeed be engaged with and feel part of modern biology, its value to society lies most strongly in its contribution to biological infrastructure but that this in no way suggests that its scholarly value should be diminished. The academic value of medical research is not reduced because of its underlying value to society. If taxonomy has declined, it has been largely through a loss of confidence in the value of its results and the lack of a collective vision: It has been said often (e.g. Ashburner, 2002) that taxonomists seem better at disputation than presenting a view that is united when communicating the need for the subject to the wider world.

Taxonomy is, of course, part of modern biology. Even a slight acquaintance with the taxonomic literature of recent years has shown that the community has adopted new techniques. Molecular methods have been used extensively for phylogenetic analysis and, more recently, the value of molecular sequences is being explored for 'bar coding' species. Certainly much funding and human effort has gone into molecular systematics, although this has been at some expense of training the next generation of organism-centred taxonomists with the capacity to examine critically all kinds of data and to synthesize the information into coherent and comprehensive results of the kind found in taxonomic monographs. Problems over the decline of this kind of taxonomy have probably been exacerbated by the molecular revolution, which is not to dismiss the value of molecular methods but rather to suggest that it is part of a greater (informatics) whole.

This scenario is dispiriting to those who believe that the special strength and value of taxonomy lies in describing the diversity of life, the word 'describing' being broadly circumscribed. But a new approach to the perception and practice of taxonomy seems to be emerging. This approach was articulated by Godfray (2002), who proposed that taxonomy should be rendered 'unitary'. By unitary was meant that the results of taxonomy for any given group of organisms should be migrated wholesale to the Internet at an authoritative site, with the content managed by an international editorial committee. Controversially, it was proposed that once a taxonomic treatment of a given taxon was accepted, following peer review, the classification and nomenclature, and presumably the taxonomic concepts subtended therein, would be fixed and changed only by future work rather than the discovery of older species names. Godfray's paper stimulated much debate and some dissent. Two further issues that require resolution are how to deal with multiple classifications of a taxon, and whether data should be 'warehoused' on a single server or made available through a distributed system via a common access system.

This contribution promotes the view that there are two complementary approaches occurring in taxonomy. One is that the most important tasks of taxonomy are those that have endured since the time that the subject emerged as a distinct discipline. These traditional activities include the description of species and other taxa. They subtend effective delimitation, identification and, more broadly, adding a wider information context (phylogeny, distribution, hosts, etc.). The other component is novel (at least relatively so) and represents the new medium of the Internet.

Although novel analytical and practical methods have been developed to the benefit (potentially or actually) of taxonomy, the graft of taxonomy goes on. By 'graft' I mean such tasks as the identification and description of new species, determining synonymy, avoiding homonymy, distinguishing species from other species and preparing keys. But by graft I do not mean to suggest any inferiority of the products. The richness and value of a good monograph surpasses most or all other subgenres of taxonomy, and the intellectual input can be every bit as demanding. New money in any science has a tendency to follow new methods and ideas. This is no different for taxonomy, but taxonomists should take care to ensure that when they seek funding they achieve outcomes of empirical as well as methodological value. We operate in an increasingly utilitarian culture and should not assume that we shall continue to gain even the level of funding for taxonomy that is currently achieved primarily for developing new methods or to produce phylogenies. Fortunately, taxonomy can be both scholarly and useful.

Schram (2004) suggested that taxonomy is a relativist rather than a realist science. Science of the realist form is aimed at falsification or verification to determine, as far as is possible, the truth. Relativist science is more fluid and concerned with organizing information to pose alternative hypotheses, and it remains in a state of constant flux as more information becomes available. The point that can be drawn from this perception is that we do not need to apologize for descriptive taxonomy because it does not fit the realist model. Perhaps too much effort has been expended to make taxonomy supposedly respectable by shifting its effort towards a more realist mode.

This contribution follows others (e.g. Godfray and Knapp, 2004) in suggesting that the emphasis for much of taxonomy in the near future is likely to be on those traditional tasks of description, clarifying and stabilizing nomenclature, and providing more effective means and protocols by which species (and other taxa) can be identified. But to maximize the value of the output, results will have to be coordinated by the taxonomic community and made more accessible than is typical at present. The Internet provides the means by which such access may be achieved. Information and communication technologies also enable taxonomists to work together in distributed teams or networks.

REVIVAL

There are several signs that descriptive taxonomy is undergoing a revival, even if, like descriptive ecology, it has been unfashionable for many years. The signs come from opinion-type papers that have been published recently (e.g. Godfray and Knapp, 2004; Wheeler, 2004), the attractiveness of the Internet as a medium on which taxonomists can post their results and make them far more accessible than the vast paper archive that exists, and the fact that funding bodies have evidently backed a number of projects involving the use of the Internet. Notably, support has come or is coming from: the US National Science Foundation with its Planetary Biodiversity Inventory programme; the British Research Councils (e.g. a BBSRC/EPSRC initiative in bioinformatics, which included a number of biodiversity as well as genomic informatics projects, and the Natural Environment Research Council [NERC] e-science initiative); The Leverhulme Trust; the Australian Biological Resources Study; and the European Commission, which, pending contract negotiation, most recently has agreed to fund a consortium of institutions to undertake a taxonomic Network of Excellence (EDIT–'The a European Distributed Institute of Taxonomy'), funded under the European Union's Framework Programme 6.

This revival seems to be being driven by the realization, which has extended beyond the taxonomic community, that describing life and providing a means of identifying species (or other taxa) is a product (or a multiplicity of products) requiring research of high quality and, furthermore, research that is of value to many user groups—from taxonomists and ecologists through to the general public. Certain ecologists (e.g. Godfray, 2002; Gotelli, 2004) have sometimes been rather better publicists for the taxonomic cause than taxonomists themselves.

I use the word 'revival' rather than 'revolution' because describing biodiversity rests on tasks and protocols that have formed the bedrock of taxonomy for as long as the discipline has existed in a formal state. This traditional background may be illustrated by the fact that much, perhaps most, taxonomy is carried out in museums and herbariums within buildings that are typically of a certain kind and form–many having been constructed in the mid-nineteenth century. Such buildings may give the impression that the activities being undertaken within them are old-fashioned. Any such perception reveals a lack of critical thought. There is no need to perceive the fabric of museums or the activities within them as obsolete. This is not to say that new, well-planned facilities are unwelcome, particularly when they improve protection for collections. Rather, it is to appreciate that the taxonomic process requires access to the considerable information legacy of the past, and that even when combined with new data, that legacy forms a large part of the current taxonomic infrastructure.

A taxonomic revival certainly does not and cannot mean just more of the same. The emergent vision for this 'new' (descriptive) taxonomy is probably best understood as one of blending a well-tested and robust system with what has already proved to be a revolutionary medium. The Internet is no longer new to early twenty-first century beings, but new Internet technologies will continue to broaden access and render it more effective.

Reviving taxonomy does not mean that its quality should be compromised. But it does mean that certain practices will change. For example, it will be possible to have high-quality taxonomy posted dynamically on the web without having to wait for a completed taxonomic revision. All taxonomy is properly regarded as work in progress, but such a mechanism makes that observation more apparent.

Any major taxonomic website will need to ensure that it carries authority without being authoritarian. And it will certainly require both the backing of the appropriate expert network of specialists and the inclusion of classifications other than the consensus. The content will have to be designed and layered so that it is accessible to multiple audiences: the peer community of taxonomists; scientists in other fields, such as ecology and conservation; journalists; teachers; policy makers seeking the best research on the taxonomy of a particular group of organisms; and the public. One great advantage of an e-taxonomy is that it can be much more image rich than is typically the case for works published on paper. An Internet-based taxonomy with content designed sensitively will be of social relevance, and we should remember that society funds much of taxonomy. Such access can hardly but improve the standing of taxonomy.

THE NEW SYSTEMATICS

In *The New Systematics* (Huxley, 1940), Julian Huxley expressed the view that the subject should become informed and engaged with species concepts based on our understanding of genetics. He hoped that through 'better liaison with … other braches of taxonomy', museum taxonomists would 'escape from the burden of routine description and naming and take full part … in the New Systematics'. Implicit in these comments was that taxonomy might at last become 'real' science–more conceptual and experimental and less descriptive. Describing species was, evidently, a 'burden' and the approach explored in the New Systematics of 1940 was the answer to helping systematics 'cope with the burden of its own data'.

That taxonomy is data rich is unquestionable. Yet far from being a burden, good quality data are the very strength of taxonomy. At least Huxley realized the magnitude of the information problem but, understandably in pre-Internet days, failed to anticipate the real work of systematics that is emerging today. Increasingly, that work is about managing an information-rich field by gathering and integrating new information using methods of data exchange available now.

Two observations suggest that we might view Huxley's burden in a more positive light. The first is that one way of perceiving taxonomy is as an information science. Information science is typically concerned with gathering, manipulating, storing, retrieving and classifying recorded information. Of course taxonomy is more besides, but this perception offers possibilities for exploring and managing the discipline's rich data source in new ways. There is no question but that computing, particularly information and communication technologies, has opened possibilities for taxonomy, notably in terms of data storage and access and in analysis. The data richness of taxonomy includes not only that which exists in the literature, but also its material infrastructure of specimens.

A second observation is that collections-rich institutions are scattered across and between nations, together with the specimens and archives they hold and the specialist staff they employ. At present there are rather few strategic alliances between institutions, although there is much communication between taxonomists from different institutions. An alternative to the currently 'fragmented' arrangement is one that encourages taxonomic resources, material and human, to be 'distributed'–that is, spread but better interlinked. This distributed model is very appropriate for taxonomy and it contrasts with one often encountered in science of having a few consolidated centres. Given the number of vulnerable collections that exist, a certain amount of consolidation is likely (and, to save threatened collections, highly desirable), but it seems unlikely that nations will easily give up their collections to centralized infrastructures (Scoble, 2003, p. 14). Considerable progress has been made to link specimen (unit)- and collection-level databases and to make their information accessible through interoperability software (e.g. BioCASE; www.biocase.org). The same is occurring at the species names level, notably in the federated system of nomenclatural databases connected to Species 2000 (www.sp2000.org). But there is much further to go.

The distributed approach to taxonomy, both in practice and perception, is increasing. A consequence of such an arrangement is that the specialist knowledge associated with collections-rich institutions is also distributed. It has always been the case that no institution has complete taxonomic coverage. What has changed is that electronic networking is now embedded in our culture, so that the many personal visits made by taxonomists to sister institutions are complemented by e-mail contact and, increasingly, by access to online databases of specimens and collection-level data (including images). Such improved communication helps reveal gaps in expertise, which should encourage better coordinated recruitment across institutions. The taxonomic community has rarely planned recruitment across institutions, but if taxonomy is to become an effective 'big science', far better coordination is essential. This issue, among many others, will be addressed in the EDIT project mentioned earlier.

The distributed nature of taxonomic resources is further underlined by the importance of the amateur community, which has contributed a substantial amount of taxonomic information to our collective knowledge base. Indeed, in some taxa most of the contribution comes from amateurs and most of the expertise lies within that community. The amateur community is even more widely spread than the professional one and will need considerable encouragement to become linked integrally to any taxonomic networks. The role of amateurs in biodiversity work in Britain (including elements of taxonomy) is being studied in the 'Amateurs as Experts' project (www.lancs.ac.uk/fss/projects/ieppp/amateurs/index.htm).

THE NEW TAXONOMY?

Few would dispute Huxley's point (1940) that taxonomy should be closely engaged with other fields of biology, for 'taxonomy stands alone, divorced from the rest of biology, at its peril' (Godfray and Knapp, 2004, p. 560). But whereas Godfray and Knapp emphasized the independence of taxonomy as a discipline in its own right and with a future particularly in its descriptive side, the tenor of Huxley's paper emphasized a need for taxonomy to become far less descriptive and much more engaged with population genetics and ecology.

Taxonomy is fragmented. Its outputs are published largely on paper, thus impeding availability to those taxonomists lacking extensive personal libraries or with access to large institutional libraries. Electronic access to published results is helping overcome the impediment, but there is much legacy literature to digitize, a resource that, unlike the situation in most of science, is of great value and relevance to taxonomy. In the new taxonomy, consolidated web revisions should allow users a single point of access to authoritative and refereed information on specific taxa. The CATE project (Creating a Taxonomic E-Science, www.cate-project.org), aims to provide consensus web-based taxonomic revisions benefiting from a peer-reviewed process built into the workflow.

Unfortunately, while migrating taxonomy to the web potentially makes information far more available to the large number of people with Internet access, it does not solve the problem of fragmentation. Fragmentation is alive and well on the web, with ever more websites for a given taxon being constructed and launched. There is no particular objection to taxonomists posting information on the web–after all one of the great strengths of the medium is the democratization of information, much of it already paid for by the taxpayer. But such information is not the desired single source of access to an authoritative revision of any particular taxon. Nor do many sites indicate the level of quality control over content; few are refereed, which is a major impediment to encouraging taxonomists to publish on the Internet.

How can the new taxonomy be achieved? Most important of all, before we become seduced by the attractions offered by cyberspace, we need to keep creating content. This includes both critical analysis and synthesis of data already in the literature and in curated collections, and in making new knowledge available. Herein lies the hard work of (monographic) taxonomy; however, it is essential since the results comprise our collective understanding of the diversity of organisms. But whereas, in the past, content was typically developed by individual effort, the Internet now provides a means by which taxonomists potentially can work in a much more coordinated way by remote linkage through information and communication technologies and by providing cyber access to their results.

The new taxonomy is likely to be ‘big’ taxonomy because to flourish it needs to be ‘big science’ (Wheeler, 2004). If the ‘big science’ label is appropriate for taxonomy it should mean better support. Fortunately, the strongly distributed nature of taxonomic resources and expertise lends itself to such an approach. Moreover, a major function of taxonomy is to describe biodiversity, providing the field with a social relevance of considerable importance. If loss of biodiversity is comparable in terms of societal concern to that of climate change, why, it was asked, are taxonomists failing to make the impact they should (Anonymous, 2004)?

WORKING TOGETHER: INDIVIDUALS, TEAMS AND NETWORKS

Big science is typically highly collaborative with its effort carefully organized, but this has not been the usual style of work for descriptive taxonomists. Some of the best taxonomy has come, and doubtless will continue to come, from solo workers (most of whom are, nevertheless, well connected with other colleagues). There is, however, no reason why a greater and more formalized level of collaboration should not be adopted by the discipline. There are examples where this is occurring, beyond molecular systematists who work in laboratories where teamwork is the norm. One of the best examples of such teamwork in descriptive taxonomy is found in the Planetary Biodiversity Inventories (mentioned earlier; also, see Knapp, this volume) with researchers focused on selected taxa. Another, which is emphasized in this chapter, is found in networks. Networks are much favoured by the European Commission in their Framework Programmes and the taxonomic community in Europe has benefited from support in the building of infrastructure for taxonomy. These projects have not, to date, been concerned directly with providing the full content found in taxonomic monographs, but they do demonstrate how taxonomy can be undertaken in a collaborative way.

The term ‘network’ refers in these EU-supported projects both to networks of connected databases and to networks of people. There are several examples of initiatives dominated by natural history museums and botanical gardens–that is, institutions with a focus on taxonomy. BioCASE, a transnational network of biological collections, provides users with access to numerous specimen- and collection-level databases across Europe. Collection-level data are gathered from 31 National Nodes to a core metadatabase, providing descriptions of what can be found in many European collections. Specimen (unit)-level data are linked to the BioCASE system directly, not through the National Nodes. Effort was spent on the BioCASE user interface to aid access. This project was pan-European and involved taxonomists interested in improving access to information in collections. Although funding for the project has ended, the infrastructure continues to be built by the connection of further databases, some through the slightly later ENBI initiative. The effectiveness of such a network to users depends on several criteria, including the number of linked databases, the effectiveness of the technical system for gaining access to the data and the quality of the data. What has also proved essential has been the underlying wish to collaborate by the project partners, the National Nodes (where these differ from the partners) and additional data providers. The BioCASE project developed from two smaller projects, also supported by the EU: BioCISE (www.bgbm.fu-berlin.de/biocise/default.htm), which dealt with collection-level data, and ENHSIN (www.nhm.ac.uk/research-curation/projects/ENHSIN/), which was concerned with unit-level (specimen) data.

Another European consortium, formed by 20 institutions holding substantial natural history collections, is SYNTHESYS (www.synthesys.info/). Besides facilitating visits of taxonomic researchers to European institutions outside their own national boundaries, SYNTHESYS includes a large networking component focused on collections with the purpose of creating a single ‘virtual’ museum service.

ENBI, the European Network of Biodiversity Information (www.enbi.info), with over 60 participants engaged formally, developed an open network of European biodiversity institutions and provides European input to the Global Biodiversity Information Facility (GBIF).

Although most of these networks are backed by consortium agreements, to function effectively all require the goodwill of the participants and an understanding of common aims within and between projects. While large projects nearly always encounter some managerial problems, the European Commission-funded projects have demonstrated that institutions across Europe have the will to work together. The ideal position for taxonomy would be to reach the stage at which the various human networks become sufficiently effective so as to become part of the thinking of taxonomists and the standard way of working. The most difficult problem to be overcome is giving credit to individuals for their input and enabling their outputs to be properly credited, citable and accepted as a proper part of performance by management.

NEW ‘BIG TAXONOMY’ INITIATIVES

The networks of people noted earlier are focused on building taxonomic infrastructure. Some projects, while contributing to that end, are partly or wholly intended to develop Internet access to more comprehensive data on taxa–notably descriptive work.

European Distributed Institute for Taxonomy

This latest EU project, which has been noted earlier, started in 2006. It is of particular importance to the new taxonomy. EDIT, the European Distributed Institute for Taxonomy, is a Network of Excellence with 27 European, North American and Russian partners. The purpose of EDIT is to reduce fragmentation in taxonomy; promote integration of taxonomic effort, particularly in leading European institutions; and promote collaboration. Two of the EDIT parcels of work ('workpackages') that are of special significance to the new taxonomy are those that deal with the promotion of web-based revisions and the construction of the Internet platform to increase the efficiency of the taxonomic work process by, for example, providing tools for key-generation and descriptions.

The plan of work for the web revisions will address four topics. The first is to describe the form of the content and its presentation. There are many commonalities in taxonomic revisionary work of any group of organisms, whether or not covered by different codes of nomenclature. Second, it is necessary to define what is needed from the Internet platform on which a web revision is to be based. This work is being carried out in close conjunction with the work package dedicated to constructing the platform. The third task is one of management. For each taxon for which a collaborative web revision is to be constructed, it is intended that an international working group is set up to run an expert network of specialists on the group. This task is easily enough stated, but it is demanding to put into operation. Classifications are nearly always subject to dispute and taxonomists hold to differing taxonomic concepts. A major requirement of any web revision will be to develop a website that represents a current, refereed consensus classification with adequate representation of competing views. This sociological component may be the most difficult part of getting web revisions produced and posted. The final area of work is to examine the several implications of publishing in the web environment. Currently, web publication of new taxa is not permitted under the codes of nomenclature because a fixed medium is required. It is likely that this position will change in the near future, but there is a clear need to have any changes safely archived: The dual use of Internet and paper is likely to be needed.

Legacy Infrastructure Network for Natural Environments (LINNE)

A major initiative currently being developed in the USA is LINNE, the Legacy Infrastructure Network for Natural Environments (www.flmnh.ufl.edu/linne/) (see Woolley, this volume). The stated goal of the project is 'to accelerate taxonomic research and improve biological collection infrastructure so that reliable information on biological diversity is available to all branches of science and society'. LINNE is at a formative stage but, if it is successful, will enable taxonomy to claim immediately big science status.

Creating an e-Taxonomy (CATE)

A project directed explicitly towards descriptive web-based taxonomy has been funded by the UK Natural Environment Research Council (NERC) as part of its e-science initiative. CATE's primary purpose is to develop a system and workflow for constructing web-based taxonomic revisions. Key outputs are two 'consensus' web revisions as demonstrators for the specialist and wider communities. The practicalities involved include setting up such a system in terms of the form of the content, wider collaborations and developing a robust and scaleable IT system to enable the content to be incorporated and made widely accessible. The demonstrator taxa selected include an animal group (Sphingidae, Hawkmoths) and a group of plants (the Araceae, Aroids). A major purpose of this project is to produce taxonomic data that will be of value to a wide user base, notably and particularly the NERC community. Besides the technical work to construct the revisions, the broader dimension and generalities of undertaking this project will be explored. These include dealing with taxa subject to different codes of nomenclature; identifying and implementing those aspects of taxonomy that can be done in a web-environment, but not on paper; dealing with refereeing mechanisms in web revisions; and providing credit to authors publishing on the web. In providing material support for this work, the NERC has signalled a strong and positive signal for the development of web-based taxonomy. The CATE system is described in more detail and the controversial issue of consensus taxonomy is explored by Godfray et al. (in press) and Scoble et al. (2007).

CHALLENGES

The road ahead will not be straightforward, for there are many matters to be resolved in the process of achieving consensus web taxonomies. Some of these are as follows:

- Gaining a consensus as to what is an authoritative classification will be difficult or in some cases impossible. A slightly different concept exists virtually every time a species or other taxon is described. Even if a broad consensus is achieved, other taxonomic concepts must be included for a web revision to be comprehensive and to avoid the risk of alienating members of the taxonomic community with different views.
- Web publication lies outside the codes of nomenclature at present. How likely is it that the codes can be changed? As standard bearers, with an underlying function to achieve nomenclatural stability, the codes are (understandably) resistant to change.
- Databases are demanding to sustain. A review of 89 bioinformatics databases (Anonymous, 2005), showed that 7 had been shut down, 44 were struggling, and only 18 were without problems; there was no information for the remaining 20. While bioinformatics databases are not equivalent to biodiversity informatics databases, similar problems of sustainability are likely to be encountered in the latter. There is surely an important role here for major taxonomic institutions to take on the responsibility of database curation besides specimen curation.
- Commitment to web taxonomy across the community inevitably will encounter resistance and lack of interest from some taxonomists. However, given the visible advantages for the discipline and for individual taxonomists, and

the level of interest and publicity in the web as a medium for the new taxonomy, we should feel reasonably confident of initial engagement by many and gradual sign-up by others. Providing ease of data entry for practicing taxonomists will be an important factor in gaining the desired engagement.

- A clear system of peer review will have to be implemented for web taxonomy. The system will need the confidence of taxonomists, but should provide users of web products with a feeling of confidence. Websites will need to be layered, so that the primary and refereed component of the web revision can be distinguished clearly from those sections that allow information to be posted prior to peer review.
- Publications and citations are important performance indicators for scientists and institutions–taxonomists included. Taxonomists will need to feel reassured that their web outputs are given adequate accreditation. There is no doubt that the Internet is being used in the publication of results in all areas of science and that at present we are in the midst of what would appear to be a radical change in the way in which scientific results are published. The senior management of institutions has an important role to play in internal accreditation of the work of their taxonomists. The issue of accreditation will be most difficult to achieve when web revisions are compiled by many individuals, which, of course, is the very model that is intended. Taxonomists are not encouraged to publish descriptions of single species, but the dynamic of a web revision will allow such descriptions to be incorporated, thus making use of the special quality of the new medium. A change in the sociology of doing taxonomy will need to be matched and, indeed, encouraged by some material changes in managerial approach. It is essential that current systems do not impede a change towards the new taxonomy.

WHAT ARE THE POTENTIAL GAINS FROM E-TAXONOMY?

To summarize, taxonomy, as is the case for other areas of science, has already benefited from the existence of the Internet (a selection of websites are categorized by Scoble, 2004). There are also a number of websites with descriptive information about taxa (e.g. Echinoidea [www.nhm.ac.uk/research-curation/projects/echinoid-directory/]; Fish [www.fishbase.org]). Most taxonomic websites, some of them very impressive, are the achievements of single authors. But, ideally, what is needed are single points of access on the web providing richly illustrated descriptions of species for every major taxon. Such sites should be comprehensive, with the flexibility to add descriptions of new species and incorporate new changes to the classification. While care should be taken not to over-complicate the information technology, the simplest solutions may not always be the best. For example, a problem with html pages is that updating them depends on the continued work of the taxonomist creating the content. Developing the content in database format is a desirable step up from the flat pages, providing greater flexibility and capacity. Provided an institution can maintain the database and the web interface for a given taxon and ensure that data can be incorporated from the wider community, this ‘warehouse’ approach to integrating distributed information (e.g. Hobern, 2005) can be effective. Alternative approaches (see Hobern, 2005) include a federated network and an indexed network. In both, data remain with the data providers, who can manage them as they wish. The federated approach is distinguished from the indexed one in that the latter allows for queries frequently issued to be handled through an index, rendering such queries easier to effect. If the disadvantage of the warehouse approach is that it may be difficult to sustain once the coordinator is no longer active, disadvantages of the federated or indexed solutions are that data quality, coming from a variety of sources, will vary and the level of coordination there will mean that the content will not be synthesized critically.

Contrary to current feelings, taxonomists have every reason to be optimistic about the future. We have the resources to do descriptive taxonomy: collections, literature, expertise (although that is probably declining) and the support of funding bodies and governments. We have a clear social need for describing biodiversity given it serious decline. If that optimism turns to despair (and for many taxonomists it has come close to that position), it will probably be the result of taxonomists failing to rise to the informatics challenge and by managers of collections-rich institutions failing to recognize what their staff are good at doing and what their institutions are exceptionally resourced to do well. Collections-rich institutions need to recover their confidence and forge ahead.

ACKNOWLEDGEMENTS

I thank Quentin Wheeler for requesting I include this chapter and organising a stimulating symposium where some of these thoughts were aired.

REFERENCES

Anonymous (2004) Ignorance is not bliss. *Nature,* 430: 385.

Anonymous (2005) Databases in peril. *Nature,* 435: 1010.

Ashburner, M. (2002) Correspondence. *Antenna,* 26: 73–74.

Godfray, H.C.J. (2002) Challenges for taxonomy. *Nature,* 417: 17–19.

Godfray, H.C.J. and Knapp, S. (2004) Introduction. In *Taxonomy for the Twenty-first Century* (eds H.C.J. Godfray and S. Knapp), *Philosophical Transactions of the Royal Society, Series B,* 359: 559–569.

Godfray, H.C.J., Clark, B.R., Kitching, I.J., Mayo, S.J. and Scoble, M.J. (in press). The web and the structure of taxonomy. *Systematic Biology.*

Gotelli, N.J. (2004) A taxonomic wish-list for community ecology. In *Taxonomy for the Twenty-first Century* (eds H.C.J. Godfray and S. Knapp), *Philosophical Transactions of the Royal Society, Series B,* 359: 585–597.

Hobern, D. (2005) Putting it all together: Technological integration at the global level, http://barcoding.si.edu/LibraryAndLaboratory/4-16_HobernPuttingItAllTogether.pdf.

Huxley, J.S. (1940) Introductory, towards the new systematics. In *The New Systematics* (ed J.S. Huxley), Oxford University Press, Oxford, pp. 1–46.

Schram, F.R. (2004) The truly new systematics–Megascience in the information age. *Hydrobiologia,* 519: 1–7.

Scoble, M.J. (2003) Introduction, Changing roles and perceptions in European natural history collections: From idiosyncrasy to infrastructure. In *ENHSIN: The European Natural History Specimen Information Network* (ed M.J. Scoble), The Natural History Museum, London, pp. 11–20. [Online at http://www.nhm.ac.uk/research-curation/projects/ENHSIN/index.html].

Scoble, M.J. (2004) Unitary or unified taxonomy? In *Taxonomy for the Twenty-first Century* (eds H.C.J. Godfray and S. Knapp), *Transactions of the Royal Society, Series B,* 359: 699–710.

Scoble, M.J., Clark, B.R., Godfray, H.C.J., Kitching, I. and Mayo, S.J. (2007) Revisionary taxonomy in a changing e-landscape. *Tijdschrift voor Entomologie*, 150: 305–317.

Wheeler, Q.D. (2004) Taxonomic triage and the poverty of phylogeny. In *Taxonomy for the Twenty-first Century* (eds H.C.J. Godfray and S. Knapp), *Transactions of the Royal Society, Series B,* 359: 571–583.

3 Taxonomy as a Team Sport

Sandra Knapp

CONTENTS

> It is a truth universally acknowledged, that a single man in possession of a good fortune must be in want of a wife.
>
> **(Opening sentence of *Pride and Prejudice* by Jane Austen, 1813)**

INTRODUCTION

Taxonomy is often characterized as a cottage industry of individuals working in isolation: the antithesis of big science as exemplified by particle physics or genome sequencing. Issues associated with new ways of working in large groups of researchers and how these relate to the use of new technologies as well as to interactions between people themselves are critical to the future of taxonomy. These human factors–both in terms of research partnerships and audiences for taxonomy–impact how individuals function within institutions; I draw on experience of large-scale taxonomic projects such as floras and NSF's more recent Planetary Biodiversity Inventory experiment to explore these issues. People are often perceived as the problem, but they must be the solution as well, or we will fail in our aim to transform taxonomy into big science that can be done as a true team sport.

Taxonomists are well aware that time is running out to describe and document life on Earth; habitat destruction and alteration continue apace and the world's population is growing at a rate that should frighten even the most complacent. For the taxonomic community, the 'truth universally acknowledged' is that taxonomic data are essential to any solution to the biodiversity crisis and thus that our discipline must carefully consider how it will become an integral part of these solutions. Another truth universally acknowledged is that taxonomy is in a parlous state. The state of taxonomic research in the UK has been investigated by two House of Lords Select Committees in the last decade (House of Lords, 1992, 2002), and the phrase 'taxonomic impediment' has become part of the vocabulary of biodiversity studies and enshrined in the Convention on Biological Diversity (see www.biodiv.org).

One of the first problems one runs up against when discussing taxonomy and its utility is that the definition of taxonomy differs depending upon your own personal or scientific perspective. For some, taxonomy is naming and systematics is the science of classification, so taxonomy is just a paper exercise and not science at all. For others, taxonomy is a part of systematics; for others, systematics is a part of taxonomy. The *Oxford English Dictionary* (Anonymous, 1971) defines taxonomy as the following: 'Classification, especially in relation to general laws or principles; that department of science, or of a particular science or subject, that relates to classification'. So, according to the *OED*, taxonomy encompasses classification. Wheeler (2004) has suggested that the 'new synthesis' in evolutionary biology came to relegate taxonomy to its second-class status as a mere part of a larger systematics, which was all about real science. Arguments about definitions can be extremely sterile, however, and can seriously impede progress in a changing world.

To define taxonomy and systematics as different, or even give them strict definitions, is an unnecessary exercise in semantics; science is best defined by looking at what it does, or its products, rather than by coining static definitions for what is a dynamic and ever changing field of endeavour. I suggest we can best examine taxonomy by looking at its spheres of endeavour and have suggested that the three principal tasks of taxonomy are description (or circumscription/delimitation), identification and phylogeny (Godfray and Knapp, 2004; Knapp, 2004). All three of these tasks help us to understand life on Earth–its diversity and relatedness. In this chapter, I wish to focus on the descriptive side of taxonomy, rather than on phylogeny or identification; this is the basis for the rest, and has long been the Cinderella of taxonomy (Godfray and Knapp, 2004).

That descriptive taxonomists are rare and a diminishing resource has been acknowledged for some time (Gaston and May, 1992), as has the fact that most taxonomists reside far from the hot spots of Earth's biodiversity. Various solutions to overcome this have been proposed: Gaston and May (1992) articulated the problem clearly, stating that 'if we found these demographic trends in a newly discovered species of lemur, we would bring specimens into a zoo and start a captive breeding programme'. Captive breeding has not yet been tried, but much else has been suggested and has been done. After the 1992 House of Lords Report the UK's Natural Environment Research Council established the NERC Taxonomy Initiative that funded many excellent phylogenetic projects and substantially contributed to the development of molecular phylogenetics. Godfray (2002) suggested that taxonomy needed to reinvent itself for the information age by becoming electronic and discarding the difficult and time-consuming past. Tautz et al. (2003) suggested that molecular biology could solve the taxonomic impediment by automating identification and species delimitation. The Global Taxonomy Initiative (GTI) hopes to revive taxonomy by highlighting its importance in the conservation of biodiversity and poverty alleviation, something with which it is hard to find fault, and via the stimulation of new worldwide partnerships (http://www.biodiv.org/programmes/cross-cutting/taxonomy/). The Consortium for the Bar-Coding of Life (CBOL, see http://barcoding.si.edu/index_detail.htm) hopes that finding universal identifiers will both solve the taxonomic impediment and provide more funds and impetus for taxonomy (Janzen et al., 2005). All of these initiatives have promise, but no one of them has yet transformed the science to reverse the downward trend.

How then do we, a relatively small and fragmented community, begin to even approach the immense task of trying to describe life on Earth before it disappears? New technologies, such as DNA bar-coding, names harvesting and automatic identification, are often touted as final solutions to the taxonomic impediment. But are they the only solutions, or only part of a more rounded and sustainable new taxonomy? People do taxonomy and other people use the results of our science; thus, any solutions or new ways of working must take into account human factors. Taxonomy as big science can learn a lot from other disciplines such as physics, strange as that may seem. Can taxonomists come together to speak with one voice while retaining the differences and dialogue that characterize a vibrant scientific discipline? If the taxonomic community's answer to this question is yes–even a qualified yes–then we as a discipline have a vibrant and intellectually stimulating future. But if the answer is no, I fear we are on the way out, just like Gaston and May's (1992) newly discovered species of lemur.

In this chapter I will not provide a blueprint for how taxonomists should work together as teams, but instead would like to explore some of the challenges for taxonomy as we settle into a new century and a new way of working. The issues I discuss here are not all that will affect taxonomists working in larger groups, but are those that have influenced me as I have worked on team projects in taxonomy. I will present ideas about the tension between doing big science and working with big questions, the influence of new technologies and then discuss the sociology of working in groups, using floristics as an example. Some of the issues discussed here are common to all taxonomy, be it done as individuals or in larger teams, but all must be acknowledged if we are to implement an approach that involves working in larger teams to transform taxonomy into an efficient 'big' science.

BIG SCIENCE OR BIG QUESTIONS?

Those debating the future of science often disagree as to whether scientists–not just taxonomists–should be doing 'big science' or addressing 'big questions'. Team science is often equated with big science, but working in teams or in large groups does not preclude working with big questions in science; in fact, it could be argued the opposite is true. The two approaches are not really in opposition, but merely different perspectives as to how to find out more about the world around us. No single scientist can alone answer a big question; even if we were to agree as to what the really big questions in science were, all science is done in conjunction with other science: Many people doing a lot of little things often achieve breakthroughs of significant import. That these breakthroughs are often attributed to a single individual or small group of individuals is a product of our societal view of achievement, not always a reflection of the context in which the breakthrough was made. This is not to deny, of course, that some individuals have indeed made groundbreaking discoveries that have changed the way we think about the world, but merely to say that great discoveries occur in a context that involves other people (scientists) and other observations. Charles Darwin's articulation of evolution by natural selection owes much to his extensive correspondence with others; without them, he might have never published (see Desmond and Moore, 1991; Browne, 2002). Networking–in the twenty-first century sense–was essential for him, as it is for most good science today.

Big science is not the mirror image of addressing 'big questions', but rather has come to be defined as doing science in large groups, each individual or research team contributing a bit to an overall solution or idea. What most think of as the 'type specimens' of big science are the particle (or high-energy) physics experiments of the mid- to late twentieth century, where the price and availability of equipment forced scientists to share. Building more than one large hadron collider (like the LHC at CERN, http://lhc.web.cern.ch/lhc/general/gen_info.htm) or linear accelerator (like SLAC at Stanford, http://www.slac.stanford.edu/) was not even an option; The Superconducting Supercollider (SSC) designed in part to find the Higgs boson was projected to cost some US$12 billion to build before it was ultimately abandoned. The team of scientists who participated in the experiment (really a series of experiments) using the Tevatron collider at Fermilab (the CDF experiment, http://www-cdf.fnal.gov/) that discovered the top quark (Yoh, 2005) was huge–more than 500 scientists from more than 60 institutions. To a certain extent, big science in physics (as we tend to define it today, but see later discussion) was driven by technology; the equipment needed to address scientific questions was too expensive or large to be built more than once. Big science begat big infrastructure. People needed to work together, not only because there was only one piece of equipment, but also because solutions to questions needed many component parts.

Genome sequencing is also what can be categorized as big science, although it is just a start to addressing a multitude of scientific questions. Many individuals, laboratories and institutions contribute to any individual sequencing effort, all compiling and checking data in order to achieve a final product: a string of annotated As, Ts, Cs and Gs (not quite a simplistic as this, but in essence the model). Genome sequencing does not depend so much on single pieces of very expensive equipment, but rather on a collaborative mindset where data are shared and checked by a large group of scientists all working towards a common, agreed upon goal. In the tomato genome sequencing project, for example, a different country has taken responsibility for each chromosome, with 10 nations contributing to the overall effort. Standards have been agreed upon, but each lab or group of labs in each country does the sequencing in their own way, and have obtained their own national funding. Checks and counterchecks are done by some labs for the benefit of all, and throughout the consortium resources are pooled and shared by all partners (Mueller et al., 2005; see also http://www.sgn.cornell.edu/help/about/tomato_sequencing.pl).

Genome sequencing as big science differs fundamentally from particle physics as big science, however, in quite a few ways. Sequencing can be done without truly expensive pieces of equipment (well, compared to particle physics anyway!), and the prices of technology in the genomics world come down yearly. Some predict that soon genome sequencing will be so routine that a $1000 cost to sequence an entire genome will be reality in the next decade (Service, 2006). As far as I know, no one is suggesting that Tevatron colliders will be in every university car park, but the world changes fast!

Neither big questions nor big science is easy to define or even to characterize; each example has subtle, but fundamental differences and cannot necessarily be easily transferred to taxonomy. Any debate over which paradigm or method to use in order to do the best science is at best sterile, but at worst a tragic waste of time and resources in debating how we should do something rather than just getting on and doing it. Scientists can be frightened to get involved in big science–even physicists! We are all trained to be individual thinkers and to critically examine every step of what we do. An analogy drawn with reference to the CDF experiments really explains big science best:

> Such a large experiment is NOT like some impersonal regimented army of hundreds of people with one leader who tells everyone what to do–but more like a cooperative of small 'families' inside a village.
>
> **(Yoh, 2005)**

The interactions of people and research groups in genomics and particle physics are really pretty much the same, despite other apparent differences.

If we look carefully at taxonomy, we can see elements of both big science and big questions in how we approach our own science. Big science in particle physics had (and still has!) big infrastructure, but so does taxonomy, in the form of the collections (of both specimens and literature) held in museums and herbaria that have been amassed over the last three-plus centuries. These collections can be considered models of the world around us (R. Lane, pers. comm.) that are refined through work with them and addition to them. Imagine trying to do taxonomy without collections: One could not check the identification of a specimen, ascertain if a plant or animal was new or already described or really assess the morphology of anything without going back in the field (a pretty expensive business) and finding it all over again. Collections are the taxonomist's equivalent of a large hadron collider or Tevatron, and they have been built at similar expense, albeit over much longer timeframes and with significantly more fragmentation (see Scoble, 2003).

Taxonomists are contributing to big scientific questions such as 'What life is there on Earth?' but perhaps too often we become narrowly focused on individual goals without looking outside our own discipline or even our own taxonomic group (i.e. nematode worms or nightshades or beetles) to see the context in which our work sits. Each taxonomist is contributing to an overall understanding of life on Earth, as each genome sequencing lab is contributing to an overall high-quality sequence, but do taxonomists see themselves as part of a big project? This depends on the taxonomist–something that I think is a real problem when it comes to how to do taxonomy more quickly and more efficiently (see later discussion). Taxonomists working on the delimitation and description of taxa tend to bridle at the assertion that their science is a service for the rest of biology, and they can get defensive about just what we do taxonomy for. Descriptive taxonomy is more than just reading a sequence of base pairs along a strip of DNA (see Lipscomb et al., 2003); the delimitation of a species is a hypothesis about the distribution of variation in nature, and it can be tested with more data or with different sorts of data (see later discussion). Species delimitations are hypothetical and predictive; they are extrapolations from a sampled subset applied to the larger, global set of organisms; and, most importantly, they are subject to test. It can be tempting for those not involved in the science of taxonomy to perceive it as a task that has to be done once and then we can move on to 'more exciting' things, when in reality few taxonomists operate in that way. Taxonomy can be big science as well as address big questions, but to date, we have not necessarily done this very well or in a very organized way. The debate for taxonomy is not whether we do big science or big questions, but rather how we can do both more effectively.

NEW TECHNOLOGIES

Technology has, throughout history, helped human beings to do things faster and more efficiently. Technological fixes are often seen as the solution to problems with how we do things, but they are not always as easy to apply as it might seem. If more taxonomy is what we need, then it must be sped up if we are to describe life on Earth and use this to overcome the biodiversity crisis (Wilson, 2002, 2004), and by doing taxonomy more rapidly we can make ourselves more relevant. Real data applied to this 'axiom' are alarmingly few; in fact, it could be argued that we actually do not need to know all species of everything in order to achieve the 2010 target of reducing the rate of biodiversity loss (see www.biodiv.org/2010). We do, however, need to know a lot more in order to even begin to measure our progress towards these targets; how can a reduction of rate of loss be measured in the absence of a baseline? The flip side of that coin is that we cannot wait long enough for a good baseline to be generated for all taxonomic groups because destruction of habitats is happening too rapidly for that. Confronted with such a Catch 22 (apologies to Joseph Heller) situation, people are often tempted to apply technological fixes–at least then we are doing something and not just twiddling our thumbs! One can think of all sorts of technological fixes to speed up descriptive taxonomy or make it easier and/or more accessible. Here, I will deal with just three of these.

The Internet

The World Wide Web, first conceived in the 1980s as a tool for sharing data by physicists at CERN (Berners-Lee, 1999) and an extension of the computer linking networks (e.g. ARPA-Net) set up in the USA in the 1970s (Abbate, 1999) has completely revolutionized the way people in today's society seek and access information of all sorts. The interconnectability and cross-linkages enabled by cybertechnology mean that, in a matter of seconds, a user can access a wide variety of information previously unavailable. Godfray (2002) was absolutely correct when he said that taxonomy was made for the Internet: Data- and illustration-rich, taxonomic work has multiple linkages and needs to be accessed by a multiplicity of users. The free access to Internet resources (until now, but things could change in the future; see Cukier, 2005) means that taxonomy on the Internet can be accessed by a far wider community than currently use primary taxonomic information. But which user community should taxonomists target with Internet taxonomy–other taxonomists, ecologists, politicians, interested amateurs, the 'general public'?

To be really useful, information on the Internet needs to be layered and targeted quite differently than the traditional scientific literature. Ideally, taxonomy on the Internet should serve all users, but that requires taxonomists to develop new skills or, better still, to develop partnerships with those who have the requisite skills in data presentation and web publishing. These partnerships are developing (see Scoble, 2004 (Chapter 2), but it takes time. Our own experience with the PBI *Solanum* project has taught us that to publish taxonomy on the Internet effectively and usably, it is not enough to just put taxonomic work in html or pdf files on a website; it is as important to use the medium to develop linkages and audiences for taxonomic work.

At the moment, all three codes of nomenclature do not allow publication of new taxa in the Internet (Knapp et al., 2004), although the International Code of Zoological Nomenclature (Ride et al., 2000) does allow publication by electronic means on CD or other reproducible media (see www.iczn.org). Discussions of electronic publishing at the 2005 International Botanical Congress in Vienna resulted in a recommendation in the code outlining the characteristics of what the botanical community would like to see in electronic publication of taxonomic novelties, but significant obstacles still need to be worked out, preferably in collaboration with publishers (Knapp et al., 2006, 2007). Descriptive taxonomy completely on the Internet and instantly accessible is not a short-term possibility, although the establishment of the registration system "ZooBank" (Polaszek, 2005, this volume) may drive change in the near future. Although registration of names will allow easy access for all users to all names, neither registration nor even publishing primary descriptions online is likely to significantly speed up the process of doing taxonomy. The time-consuming stages of the taxonomic process are usually the ones of careful examination of specimens, often in many different museums around the world (but see later discussion), not necessarily the writing and submitting of the new descriptions themselves. But what could be significantly sped up through the medium of the Internet is the publication process itself, although this does not necessarily mean electronic publication must happen. The average time from submission to publication for a botanical novelty is anywhere from 6 months to a year, far too long! There would be an outcry in the genomics world should this sort of timeframe be posited for publication; we as a community need to demand better service from our publishers (and via them our reviewers) or find different ways of publishing.

The Internet is a tool that taxonomists can use to our great advantage–a point upon which most taxonomists can agree. How we do this taxonomy on the Internet is another matter; the state of play today is in such rapid development that what we see as ideal today may seem clunky and outmoded in a decade. Godfray (2002) suggested that it was time for us to use the Internet to bring about the second bioinformatics revolution in taxonomy–the first having been that begun by Linnaeus with the establishment of binominal nomenclature. It is important to remember that Linnaean nomenclature (the currently used form of scientific names) was not a method authoritatively imposed from a centre, but was in a way the product of natural selection: It worked best, so it prevailed in the end. The current multiplicity of approaches to taxonomy online is healthy and will ultimately lead to new developments and a real Internet taxonomy, rather than just the taxonomy of the printed page put up on the web.

Virtual Access to Specimens

One extremely important result of the electronic provision of information is that a taxonomist in London, for example, can access easily information on specimens in Capetown or St. Petersburg, and vice versa–provided, of course, that the institutions holding these specimens make that information available. It is clear to most practicing taxonomists that an image of a specimen is the next best thing to the specimen itself, rather than just the data associated with that specimen. Electronic sharing of specimen and name data through databases online (in botany, invaluable resources for all taxonomists are TROPICOS, www.tropicos.org, and IPNI, www.ipni.org; see Nic Lughadha, 2004) but to really *do* taxonomy using these electronic resources, we need to be able to see the specimen itself. Here, technology is advancing at a dizzying rate; 10 years ago, producing and serving high-quality images on the Internet was difficult, but today many institutions and projects are exploring different ways of imaging and presenting specimens of all sorts. Herbarium specimens have provided much of the initial data on how to serve images online, but the capture of all specimen data can still be time consuming, although automated solutions are being explored (see http://www.peabody.yale.edu/collections/bot/botcurr_db.html).

Static images are fine, but ones that can be measured and compared are even better; for this, herbarium sheets are ideal; as two-dimensional objects, they are perfect for this sort of use. Publicly available tools are being developed in various institutions (e.g. in Berlin–see http://ww2.bgbm.fu-berlin.de/herbarium/default.cfm, or through the African Plants Initiative project of Aluka–see Smith, 2004, and http://www.ithaka.org/aluka/content.htm#plants) that allow users to accurately measure plant parts and to zoom into particular characters in real detail. Structures requiring dissection, however, such as long-tubular flowers or closed buds, still require the actual object; specimens themselves are still, and will forever be, essential. Some zoological specimens, insects on pins, whole worms or even skeletons can be more difficult both to image and to present on screen; a user always want to see the other side, or a different angle. Interactive viewing of specimens using high-resolution, remotely operable, digital microscopes (such as those currently being installed in London, Paris and Washington, D.C.) could provide the answer, but a person is still needed to load the specimen into the machine; technology still needs the human touch.

The electronic public domain is becoming increasingly data rich with information about and images of specimens of all kinds. This has helped taxonomists to do their science more efficiently, as the click of a mouse can let one know whether it is necessary to fly to Berlin or New York in order to see the type of some potential synonym. Literature online and publicly available through open source access is also a resource of incalculable benefit to today's taxonomists (see http://www.botanicus.org and http://www.biodiversitylibrary.org). The 'biodiversity commons' (see Moritz, 2002) for the equitable access to public domain digital information not only will benefit today's taxonomists, but also will speed the development and training of taxonomists in the biodiverse regions of the world where they are needed most and are the rarest (Gaston and May, 1992).

Somewhat perversely, however, the availability of all this information comes at a cost: A single scientist still needs to find the time to digest it all to come to a taxonomic decision. The ever increasing number of specimens also means that to come to a decision, a scientist simply has more material (specimens) to examine than did a taxonomist from the previous generation. Linnaeus (1753) saw four 'specimens' when he described *Solanum lycopersicum*, the cultivated tomato; while in the course of preparing a monograph of the group today (Peralta et al., 2008), we examined and databased almost 3000. No matter how much we put on the web or make easily available by other means, time remains a constraint that technology may not be able to overcome completely.

DNA Sequencing

It has been suggested that one way to remove the taxonomic impediment would be to convert to a taxonomy based on DNA sequences, rather than one based on morphology (Tautz et al., 2002, 2003). The pros and cons of 'DNA taxonomy' have been debated and contested intensely elsewhere (see Lipscomb et al., 2003; Mallet and Wilmott, 2003; Seberg et al., 2003; papers in Savolainen et al., 2005), and I will not review them in detail here. DNA sequences, whether of the currently identified 'barcode' (see Hebert et al., 2003) regions (COI [the mitochrondial gene cytochrome oxidase 1]) or small subunit ribosomal genes in animals or various candidates in plants (see Chase et al., 2005; Rubinoff et al., 2005; Chase et al., 2007), have proved extremely useful in working out the taxonomy of groups such as nematodes (Blaxter, 2004; Blaxter et al., 2005) or bacteria (Oren, 2004) or sibling species of mosquitoes (Krzywinski and Besansky, 2003), where morphology is either hard to see (due to size), not variable or just plain difficult. DNA barcodes as identification tools have an obvious utility and, with the costs of sequencing going down all the time, will soon be in reach of most biologists.

But the use of a barcode depends upon having a reference set against which to compare the unknown–in effect, a known taxonomy (see Janzen et al., 2005). Raven (2004) put it most eloquently:

> Finally, nothing will substitute for the activities of the field naturalist. No matter how much we speak about instant identification through DNA analysis, hand-held keys or other modern approaches, unless there are very many people who can recognize organisms, find them, go into the field and find them again, whether they be in the tropical moist forests of Congo or the chalk grasslands on the South Downs of England, nothing will work.

Some have suggested that the delimitation of taxa (i.e. true DNA taxonomy) can be done using just DNA barcodes (e.g. Brower, 1999; Pons et al., 2006), but this can be seen as a bit like defining taxa based only upon the setae on the last tarsal segment of the hind leg or on the shape of the calyx lobes. If species are hypotheses about the distribution of variation in nature (Darwin, 1859; Lipscomb et al., 2003), then multiple data-sets–of both characters such as DNA sequences of various sorts or morphology and specimens themselves–are necessary to test these hypotheses, the more data the better (Seberg et al., 2003; DeSalle et al., 2005; Seberg and Petersen, in press). It may in fact be that finding all the species (as organisms in the field!) is our rate-limiting step (May, 2004), and no matter how fast we can generate DNA sequence, we still will not be able to find organisms efficiently enough.

Whether or not barcoding life will solve all our taxonomic problems and remove once and for all the taxonomic impediment, the data being generated as part of the Consortium for the Barcoding of Life initiative (CBOL; see http://barcoding.si.edu/index_detail.htm) will be very useful for our and future generations of taxonomists. The careful vouchering of taxa from which sequences are obtained (records of which are planned to be stored in the Barcode of Life database BOLD (http://www.barcodeoflife.org/) means that the data from this initiative will be subject to repeatability and test; they will be usable for future science as well as for helping to solve some of today's problems in difficult taxonomic groups.

WORKING IN GROUPS

Taxonomy has been characterized by some as a cottage industry, with little scientists, each beavering away in his or her own corner and with his or her arcana, rarely coming out to greet the world at large. It could be argued that ALL science is a cottage industry (as compared to global finance or steel mining), but that sometimes the cottages are closer together, or perhaps even in terraces. Science is all about individual insight and effort, even when we work in large teams. Taxonomists have been working in large groups for a long time, but this teamwork is often overlooked in examinations of taxonomy, even by taxonomists themselves. Here I am not referring to the large groups of people seen in the departmental photographs of major museums, but instead use as an example the botanical discipline of floristics.

Floras–or the compilation of the taxonomy of all the plants of a particular region–have a long and distinguished tradition in botany. Many different models for how to do a flora exist; the first taxonomic work to use binomial nomenclature, *Species Plantarum* (Linnaeus, 1753), was essentially a flora of the world. *Species Plantarum* was a single-authored work (although Linnaeus had help from many others), and floras through the ages have been done by individuals; by groups of taxonomists in a single institution; and, more recently, by large, international teams of taxonomists from many different institutions, looking a bit like big science after all! *Flora Mesoamericana*, with which I have been involved for over a decade (see http://www.mobot.org/mobot/fm/), is a multi-author work, in which treatments of families and genera are contributed by experts in those taxa. Over 400 taxonomists contribute to the project, which will ultimately be a guide to the approximately 16,000 species of flowering plants and ferns of southern Mexico and the Central American republics. As in the big science of high-energy particle physics, each expert contributes his or her group to an overall publication, using the infrastructure provided by the world's herbaria.

Floras tend not to be recognized as big science, in part because they tend to take a very long time (see later discussion). The 10 volumes of *Flora Mesoamericana* will take some three decades to finish (the project began in 1980) and the *Flora of*

Tropical East Africa has been ongoing since before the Second World War (Turrill and Milne-Redhead, 1952) but has been recently revitalized (see http://www.rbgkew.org.uk/herbarium/ftea/fteaindx.html) and has outlasted the British Empire! At the time FTEA was begun, for example, it was understood that these things took time ('the preparation of the Flora will take many years'–Turrill and Milne-Redhead, 1952), but in the twenty-first century we have developed a sense of urgency as well. Many factors contribute to the rate of work done on these projects; here I will explore just four of these in the context of taxonomy as a team sport: (1) standards; (2) lack of experts and (3) deadlines; and (4) long-term funding for long-term projects.

Setting standards is always difficult when people work together in groups, especially when no 'industry' standards exist. Taxonomic descriptions are as variable as taxonomists, and when all are to be collated into a single work, consistency is necessary. Parallel descriptions make the work more useful for the end-user, but as a flora includes many taxonomic groups, completely parallel descriptions are impossibly convoluted across such a diversity of taxa: Imagine if we were to try to combine plants with animals! Excessively strict standards can also slow work down for little extra gain; pragmatism is sometimes an uneasy balance between efficiency and perfection. In printed publications, character order is important, and a user needs to know that in every description he or she can find the shape of the leaves first, then the color, then the texture, etc. But as electronic searching and data mining become more sophisticated and ontologies mature (controlled vocabularies of characters and their relationships; see http://www.plantontology.org/) character order may become less important, but getting all the key characters in a standard form will still be essential.

The different terminology for different taxonomic groups can cause problems for large projects; the detailed terminology for plant hairs, for example, is different in almost every family of flowering plants. An editor is no one's friend as he or she tries to make things consistent; the more people involved, the more arguments one has. If we really want to create a usable, global taxonomy, we will have to live with some imperfections and perhaps less than ultragold standards–something at all may have to be better than perfection. What is critically important are standards; if data are standardized then automated compilers can use them, and computers, as well as people, can 'talk' about taxonomic data. GBIF and GenBank are both good examples of open source repositories of relatively standardized data that have revolutionized research.

Even in a taxonomic group as well known and intensively studied as flowering plants, there are whole families for which there are no experts in particular regions, or no experts at all. For flora projects using expert taxonomists to contribute treatments that represent the most up-to-date and accurate taxonomic view, this means some families must either be left out (surely not even an option if what we want is a complete treatment of all the taxa in the region) or done by non-experts. In today's age of increasing specialization, many taxonomists are not willing to work outside their group of speciality; gone are the days of Carl Linnaeus when botanists were able to undertake whole floras on their own. As the taxonomic workforce diminishes (see Gaston and May, 1992), more and more holes of non-knowledge will appear, making a truly global view difficult. This phenomenon is not unique to floristics or to botany; the uneven distribution of scientists across taxa is even more glaring when looking at biodiversity as a whole.

Deadlines are a fact of life. In a multi-author flora, however, they are often perceived as someone else's (usually the editor's) problem. Taxonomic accounts in floras are often not perceived by authors as high priority, or as something that will help their career. Publication of a floristic account is not 'high impact' (see also later discussion) and thus often slips to the bottom of the always increasing pile of tasks with which today's taxonomist is faced. Contributors to floras are usually 'volunteers' and so find it difficult, completely understandably, to prioritize someone else's project. In trying to do taxonomy as a team sport, in large groups of interacting and contributing individuals, we will encounter the same problems unless we can overcome the institutional and individual barriers to such efforts.

Fortunately for floristics and for the users of these data, herbaria and botanic gardens have supported this endeavour for generations. Long-term projects need a long-term financial base, not as a go-ahead to do things at a leisurely pace, but to allow the stability that is necessary to create infrastructures and mechanisms to get such long-term work done. The ephemeral nature of funding in today's science culture, with 3-year projects the norm, means that it can be hard to even think about embarking on a decadal project that may have the rug pulled out from under it at any time. It is imperative that the big natural history institutions (here I include herbaria and botanic gardens) continue to support the long-term taxonomic projects that will constitute the new taxonomy and will ultimately be of greatest service to society at large. Godfray (2002) suggests that museums are the logical place for new taxonomies to reside, as part and parcel of these institutions' increasing roles as guardians of information. This will require two things: first, that the institutions themselves step up to the plate (to use a baseball analogy) and take on this task and, second, underpinning the first, that the governments and private foundations funding these institutions acknowledge this role as important and support them accordingly, both financially and morally.

This is not to say, however, that coffers must be bottomless. The luxury of a fully funded institution where no scientists needed to seek outside funds from grant agencies is a pipe dream, and in the long run would not be good for the science. So how should long-term science be funded? In the decades after the Second World War it became clear that to do the big science necessary to advance high-energy particle physics, long-term funding was needed. The solution arrived at in that community was to establish advisory boards as part of the funding process; these boards advised the US government departments funding the research as to the balance between short-term and long-term funding. Scientific advisory panels operate in most if not all funding agencies, government and private alike, but it was the explicit articulation of a need to balance long- and shorter term funding that enabled physics to develop the big science model of working in large groups over long periods (E.A. Knapp, pers. comm.). The National Science Foundation's Planetary Biodiversity Inventory programme is a step in the right direction (see Page, this volume) and the extension of these awards (first begun in 2003; see http://www.nsf.gov/pubs/2006/nsf06500/nsf06500.htm) for a further decade will be a model for others in how to fund the new taxonomy.

Some zoologists have said to me that floras are a waste of time because they treat only the plants of a particular region and do not take a global view. Nothing can be further from the truth. The Convention on Biological Diversity (CBD) is implemented on a national and regional scale, although we worry about it on a global scale. Floras allow local biologists to identify and monitor all the plants in their own areas relatively easily, rather than having to wade through huge global treatments of just a few groups. Training of taxonomists occurs in institutions that exist in nations where smaller sets of organisms can be easier to learn with than the immensity of global biodiversity (see Knapp et al., 2001, for a discussion of floristics and taxonomic training). Floras are a good model for how taxonomists can work in large groups–things that have gone right we can emulate and things that go wrong we can try to avoid–as we attempt to move forward and create a new taxonomy that not only moves more quickly, but also involves more people working together towards a common goal.

PUBLISHING AND ASSESSMENT

Science–including taxonomy–is assessed through publication, in general through peer review in 'reputable' journals. Results and ideas made available in the public domain are available for test and confirmation; scientists in general are suspicious of results not published in the peer-reviewed literature (e.g. the furor over GM potatoes generated by Arpad Pusztai, http://news.bbc.co.uk/1/hi/sci/tech/474978.stm). But as science grows, so does the number of publications–like specimens; the sheer number of publications an individual scientist 'needs' to read is enormous: almost 1.5 million peer-reviewed scientific articles are published every year (Mabe and Amin, 2001). Any given discipline, of course, has fewer than this, but the trend is clear. In this context the ISI (Institute for Scientific Information, started by Eugene Garfield and now owned by The Thomson Corporation) citation indices were established as a bibliographic system for cross-referencing and linking the increasing numbers of scientific publications (see Dong et al., 2005).

Part of the current implementation of the ISI Citation indices is the Impact Factor (capitalized here intentionally), a quantitative measure for assessing both individual papers and scientific journals. Increasingly, the Impact Factors are being used as a quick and easy quantitative evaluation tool for individual scientists and scientific institutions (Lawrence, 2003), something for which they were not designed. A great deal has been written about the politics of publication: the problems with the use of scientometrics as sole measures of 'goodness' of science (Colquhoun, 2003; Dong et al., 2005), statistical problems with the compilation of these indices (Adam, 2002; Anonymous, 2002; Dong et al., 2005), differences between the biomedical sciences and less 'brutal' fields (Törnqvist, 2002) and problems associated with comparing between disciplines using these simple measures (see Adams, 2005). The true impact (read 'importance') of a paper may have little to do with its or the journal's Impact Factor: An essential part of publication is choosing your audience well.

Taxonomy as a discipline does not even feature in the ISI Web of Knowledge; the closest we get is evolutionary biology, or bits of entomology, zoology or agriculture. We could emphasize the half-life measure or could try to change the way the indices are calculated in order to benefit ourselves and our discipline, but that is probably a losing battle. I feel, however, that this is not worth worrying about. What is more important is that we think about how we can achieve impact (*real* impact) with taxonomic publications–how we can make taxonomy change the world. Many aspects of publication impact the future of taxonomy as a team sport, but here I would like to focus on three of these: (1) citation, (2) publication on the Internet (see also preceding discussion) and (3) publishing in groups.

Methods papers have very high citation rates; can you think of a plant molecular phylogenetics paper that has not cited Doyle and Doyle's (1987) CTAB method for DNA extraction? Who, however, cites the original reference for the t-test? Some knowledge, when in the public domain long enough, becomes common knowledge and not cited. Taxonomic names belong to this category. Some have advocated the citation of the original paper where the species was described (so all papers on *Drosophila melanogaster*, for example, would cite Meigen, 1830, and all papers on *Solanum lycopersicum* would cite Linnaeus, 1753) or the latest taxonomic revision every time a scientific name is used. Rather than increasing the visibility of taxonomic work and literature, this would have the opposite effect. Forcing someone to notice you by these means is a sure way of getting ignored; a better strategy would be to become indispensable. Insisting on citation is something editors of journals would have to do simultaneously; it cannot be imposed by a community on the rest of science.

So how to become indispensable? Godfray (2002) has suggested that web taxonomy is one way to achieve this. But does web publication 'count' in today's bean-counting world of scientific assessment? It is essential that it does; the Internet is critical to the survival of taxonomy as a discipline, both by its ubiquity and its ability to connect us. Let us suppose that in a few years' time taxonomists and publishers have come together and devised a system for the electronic publication of scientific names that overcomes current difficulties (see Knapp et al., 2006, 2007) and that the codes of nomenclature have been modified accordingly. Then, not only taxonomic compilations (e.g. *Solanaceae Source*, our own PBI web revision at http://www.nhm.ac.uk/solanaceaesource), but also taxonomic novelties could all be available online and instantly, meaning that working as a large team revising a taxonomic group could really be done as a group activity. Electronic publication could really speed up the process of publishing taxonomy (see also preceding discussion), thus making it available to our users when it is needed, not several years later. For this to happen, however, institutions need to recognize, support and even help in the push to implement peer-reviewed taxonomic publication via electronic means, especially on the Internet.

Because publication is such an important part of today's scientific assessment, authorship is critical to progress through a career in science. If we are going to consider doing taxonomy in large teams, as big science, then we also need to consider the implications for publication and its impact on careers. In general, taxonomists do not publish large multi-author papers (the Angiosperm Phylogeny Group's output is an exception; APG, 2003), while our impression of big science is that it means papers with hundreds of authors. Comparison of a top journal in each of three fields–taxonomy (*Systematic Biology*), physics (*Physical Review Letters*) and general biology with a genomic slant (*Nature*) representing taxonomy and two disciplines

TABLE 3.1
Comparison of Author Numbers in Journals across Systematic Biology, General Biology (Molecular Biology) and Physics[a]

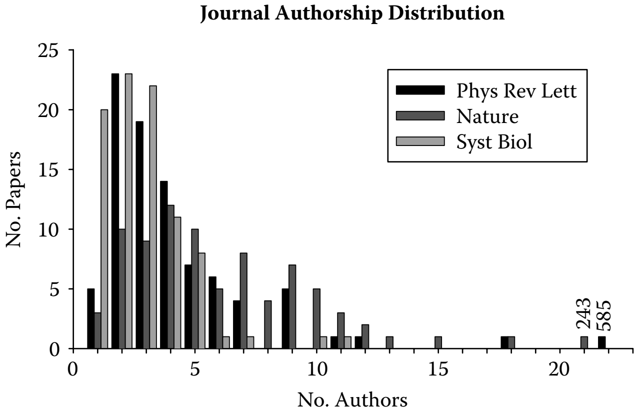
Journal	N	Range	Mean	Median	Mode
Systematic Biology[b]	88	1–11	2.84	3	2
Nature[c]	83	1–243	8.84	5	4
Physical Review Letters[d]	87	1–585	10.73	3	2

[a] Journals selected are those that publish primary research articles, not reviews. Only research articles (excluding commentaries, book reviews, etc.) were scored. See also Figure 3.1 for graphical representation of the data. Impact factors taken from ISI Web of Knowledge$^{\text{SM}}$–Journal Citation Reports (http://portal.isiknowledge.com/).

[b] *Systematic Biology,* 54 (1–4), 2005; 53: 1–6, 2004. IF = 10.2.

[c] *Nature,* 436 (14 Jul, 21 Jul, 4 Aug, 11 Aug), 2005. IF = 32.2.

[d] *Physical Review Letters,* 98 (7, 8), 2005. IF = 7.2.



Journal Authorship Distribution

FIGURE 3.1 Histogram of author numbers versus number of papers in the three journals analysed. The x-axis breaks at 20 to accommodate the two outlying papers from *Nature* (243 authors) and *Physical Review Letters* (585 authors). All three distributions are significantly different ($\chi^2 = 62.78$, df 14, $P < 0.001$). The distributions of *Physical Review Letters* and *Systematic Biology*, although having the same median and mode (see Table 3.1), are significantly different ($\chi^2 = 20.64$, df 7, $P < 0.01$), with *Systematic Biology* having more single-author papers.

we often think of as being big science–shows that taxonomy has significantly more papers with few authors, as compared to the others (Table 3.1, Figure 3.1).

The range of author numbers for all three journals (when the outliers were removed) was 1–18, but *Nature* has a significantly different distribution of author numbers (see Figure 3.1), with the peak at higher author numbers. *Physical Review Letters* and *Systematic Biology*, while having the same mode and median, do have different distributions ($\chi^2 = 20.64$, df 7, $P < 0.01$) due to a preponderance of papers with one or two authors in *Systematic Biology* (see Figure 3.1). Descriptive taxonomy may be even more skewed towards single authorship, as ideas about species delimitation or identity are often the result of one person's work. If we do begin to do taxonomy in groups, how will this change? Perhaps we will continue to have single-authored species (as an example, see species treatments in *Flora Mesoamericana* online W3FM, http://www.mobot.org/MOBOT/fm/) in multi-authored papers or treatments, be they electronic or paper.

Publications with large author numbers or by consortia have their own problems. Consortium-authored papers have been undercounted for the citation indices (Anonymous, 2002) and credit can be seen to be diluted for any individual author (except, perhaps, for the first, whose name might stay in people's minds) on papers with hundreds of authors. Evaluation of postdoctoral researchers or graduate students participating in big science can be complicated by large author numbers because employers tend to rely more heavily on personal recommendations and letters of reference–potentially fertile ground for sycophancy or even misconduct (see Martinson et al., 2005). Large projects, like those done in the big science of physics or genomics and perhaps the new taxonomy, need to be carefully constructed so that young scientists are able to claim credit for work they have done in a way that allows them to advance in their own careers. Although Lawrence (2003) was referring to the politics of where to publish, his words are worth our heeding as we attempt to create a new taxonomy that involves working in groups in very different ways: 'We cannot expect younger scientists to endanger their future by making sacrifices for the common good, at least not before we do'.

CHALLENGES FOR A TEAM TAXONOMY

In this chapter, I have highlighted some of the issues that have had resonance with me as I have embarked upon attempting to do taxonomy as a team sport, both through Flora Mesoamericana and through the Planetary Biodiversity Inventory project 'PBI *Solanum*: A Worldwide Treatment'. The perspective I have taken, therefore, is quite personal. Other taxonomists will have encountered other issues and see other fertile avenues for investigation; there are no formulae or easy fixes as we transform taxonomy into an efficient 'big' science for the twenty-first century to rapidly advance and make available to science and society reliable taxonomic knowledge. Nor do these issues pertain solely to working in teams; many of them are equally applicable to a taxonomist working in isolation, but all will need to be addressed if team working is to become the new reality. In closing I would like to highlight a few of the big challenges we face if we decide to work in larger groups–in true teams–to attain our goals:

- Harness technology. Godfray (2002) was right: Taxonomy is made for the Internet. Our use of the medium is increasing daily; the first place most botanists turn to in order to find bibliographic details of plant names is IPNI (http://www.ipni.org)–an indispensable Internet resource. The linkages and interconnectivity afforded by the electronic environment are golden opportunities. Centralization therefore might not be a good idea; there may be a period when many different systems are tried out until we settle on one that works best. (It happened in nomenclature three centuries ago, why not in the rest of taxonomy today?)

- Develop standards. Linkages to build groups are vital for working in teams, but these are facilitated by standards in data. GenBank and GBIF are examples of resources both providing and allowing contribution to standardized data; the fact that they are open source facilitates collaborative working.
- Develop constituencies. The world needs taxonomy; many communities just do not know it yet. It is far better to spend time making links with another scientific community, be it ecology or genomics or microbiology, than to complain about being misunderstood. People everywhere are fascinated by diversity, and as the genomics age comes of age, those communities in particular will need taxonomic knowledge in order to put their work in context. The linkages developed between the PBI *Solanum* project and the genomics have led to a larger audience interested in the results of taxonomy, thus raising the visibility of taxonomic work in general.
- Individual commitments to common goals. Taxonomists need to stop thinking about 'my' group or 'my' project and begin to think about 'our' revision or 'our' project. Addressing accreditation issues for taxonomic work may help to shift this. Common goals do not mean we all have to say and do exactly the same thing; diversity is, after all, at the core of our science. But in order for policy makers or big funding agencies to take us seriously we need to have a common vision–a big plan, not a robotic repetition of the same words.
- Institutional commitments. If taxonomy is to develop a vibrant new way of doing the science and making it accessible, institutions, particularly the big museums and botanic gardens, need to back the commitment of individual scientists with institutional commitment to overarching goals. Descriptive taxonomy needs support–it has been a Cinderella too long–and where better than large taxonomic institutions to be the global centres for repositories of taxonomic information?
- Funding patterns. Short-term funding is good for some science, but can damage long-term projects. In the taxonomic community, we need to begin to influence funding agencies to balance short- and long-term funding (and we are not alone in this: think of long term ecological monitoring). The recent renewal of funding (http://www.nsf.gov/pubs/2006/nsf06500/nsf06500.htm) for the NSF's Planetary Biodiversity Inventory Programme is a sign that this is happening, at least in the USA. The balance between inherent interest and tractability is always tricky, but a balance needs to be articulated, as was done in the physics community in the decades following the Second World War.
- Manage division of effort coupled with unification of purpose–probably the most difficult of all the challenges we face. It is easy to just plough ahead and do what we all do best: taxonomy on our own. But reality checks on the big plan or the global vision are critical. A framework in which to position and put in context our own work, whether done as a group or as individuals,

will do wonders to focus the mind. It is important to do taxonomy for a reason; most of us are not taxonomists doing work on our particular group 'because it is what I do', but rather we are doing taxonomy because we want to understand the diversity in the group or the morphology or even the DNA sequence variation. We just have not managed to articulate that very well on a community level.

- Above all, set our own challenging agenda, rather than just moan about the state we are in (but see Fox, 2004, for moaning as a group-building activity), with realistic and relevant goals (Godfray, 2002). These goals will not be one size fits all, but will be varied and diffuse at the beginning; however, if we, the taxonomic community, can all discover a sense of common purpose, then we will be on the road to creating a new culture of doing taxonomy.

People (i.e. taxonomists) are usually seen by critics as part of the problem with taxonomy, which may or may not be true. In order for us to truly do taxonomy as a team sport, working together in larger groups with a common purpose, people must necessarily be part of the solution. It is, after all, people who interact in large groups, and it is people who do science–who come up with ideas, test hypotheses and make new discoveries. Better ways of working together will help us achieve our goals and will help revitalize our science for this century when it is needed more than ever.

ACKNOWLEDGEMENTS

My thanks to Quentin Wheeler, who asked me to participate in the New Taxonomy Symposium in Cardiff, Wales, and the Systematics Association for hosting the symposium; Ed Knapp, whose patience with endless questions about high-energy physics was amazing; Jim Mallet, who prepared the figure and helped this mathematical illiterate with statistics; Dave Grimaldi, who helped with *Drosophila* details; Dani Zamir, whose invitation to cochair the Solanaceae Genomics Network has opened a whole new world to me; the National Science Foundation for funding through the Planetary Biodiversity Inventory programme, award DEB-0316614 'PBI *Solanum*–a worldwide treatment'; two reviewers for useful comments on the manuscript; and my colleagues at the Natural History Museum, whose intellectual company has made me think of taxonomy in a variety of different ways. Their input has been essential; all errors and misinterpretations, however, are my own.

REFERENCES

Abbate, J. (1999) *Inventing the Internet (inside Technology),* The MIT Press, Boston, 272 pp.

Adam, D. (2002) The counting house. *Nature,* 415: 726–729.

Anonymous (1971) *The Compact Edition of the Oxford English Dictionary,* Oxford University Press, Oxford, 4116 pp.

Anonymous (2002) Errors in citation statistics (editorial). *Nature,* 415: 101.

Angiosperm Phylogeny Group (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society,* 141: 399–436.

Berners-Lee, T. (1999) *Weaving the Web: The Original Design and the Ultimate Destiny of the World Wide Web,* Harper–Collins Inc., New York, 256 pp.

Blaxter, M. (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London, Series B*, 359: 669–679.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R. and Abebe, E. (2005) Defining operation taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society of London, Series B*, 360: 1935–1943.

Brower, A.V.Z. (1999) Delimitation of phylogenetic species with DNA sequences: A critique of Davis and Nixon's population aggregation analysis. *Systematic Biology,* 48: 199–213.

Browne, J. (2002) *Charles Darwin: The Power of Place,* Jonathan Cape, London, 591 pp.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., van den Berg, C., Madriñán, S., Petersen, G., Seberg, O., Jørgsensen, 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. and Wilkinson, M. (2007) A proposal for a standardized protocol to barcode all land plants. *Taxon,* 56: 295–299.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.P., Haidar, N. and Savolainen, V.P. (2005) Land plants and DNA barcodes: Short-term and long-term goals. *Philosophical Transactions of the Royal Society of London, Series B,* 360: 1889–1895.

Colquhoun, D. (2003) Challenging the tyranny of impact factors. *Nature,* 423: 479.

Cukier, K.N. (2005) Who will control the Internet? Washington battles the world, *Foreign Affairs,* 84(6): 7–13, http://www.foreignaffairs.org/20051101facomment84602/kenneth-neil-cukier/who-will-control-the-internet.html.

Darwin, C.D. (1859) *On the Origin of Species by Means of Natural Selection.* John Murray & Sons, London, 497 pp.

DeSalle, R., Egan, M.G. and Siddall, M. (2005) The unholy trinity: Taxonomy, species delimitation and barcoding. *Philosophical Transactions of the Royal Society of London, Series B*, 360: 1905–1916.

Desmond, A. and Moore, J. (1991) *Darwin,* Michael Joseph, London, 808 pp.

Dong, P., Loh, M. and Mondry, A. (2005) The 'impact factor' revisited. *Biomedical Digital Libraries,* 2, doi:10.1186/1742-5581-2-7.

Doyle, J.J. and Doyle, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemistry Bulletin,* 19: 11–15.

Fox, K. (2004) *Watching the English. The Hidden Rules of English Behaviour,* Hodder, London, 424 pp.

Gaston, K.J. and May, R.M. (1992) Taxonomy of taxonomists. *Nature,* 356: 281–282.

Godfray, H.C.J. (2002) Challenges for taxonomy. *Nature,* 417: 17–19.

Godfray, H.C.J. and Knapp, S. (2004) Introduction. *Philosophical Transactions of the Royal Society of London, Series B*, 359: 559–569.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society, London, Series B,* 270: 313–321.

House of Lords Select Committee on Science and Technology (1992) *Systematic Biology Research: Report*, HMSO, London, 169 pp.

House of Lords Select Committee on Science and Technology (2002) *What on Earth? The Threat to the Science Underpinning Conservation*, HMSO, London, 48 pp.

Janzen, D.H., Hajibabaei, M., Burns, J.M., Hallwachs, W., Remegio, E. and Hebert, P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society of London, series B*, 360: 1835–1845.

Knapp, S. (2004) Systematics: A science of three parts. *NHM,* 4: 10–13.

Knapp, S., Davidse, G. and Sousa, S.M. (2001) Proyectos floristicos hoy y mañana: Su importancia en la sistemática y la conservación, in *Enfoques Contemporaneos Para el Estudio de la Biodiversidad* (eds H.M. Hernandez, A.N. Garcia Aldrete, F. Alvarez and M. Ulloa), Instituto de Biología, UNAM, México, pp. 331–358.

Knapp, S., Lamas, G., Nic Lughadha, E. and Novarino, G. (2004) Stability or stasis in the names of organisms: The evolving codes of nomenclature. *Philosophical Transactions of the Royal Society, Series B,* 359: 611–622.

Knapp, S., Polaszek, A. and Watson, M. (2007). Spreading the word. *Nature,* 446: 261–262.

Knapp, S., Wilson, K. and Watson, M. (2006) Letter to the editor [electronic publication]. *Taxon,* 55: 2–3.

Krzywinski, J. and Besansky, N.J. (2003) Molecular systematics of *Anopheles*: From subgenera to subpopulations. *Annual Review of Entomology,* 48: 111–139.

Lawrence, P.A. (2003) The politics of publication. *Nature,* 422: 259–261.

Linnaeus, C. (1753) *Species Plantarum.* Laurentii Salvii, Stockholm, 1200 pp.

Lipscomb, D., Platnick, N. and Wheeler, Q. (2003) The intellectual content of taxonomy: A comment on DNA-taxonomy. *Trends in Ecology and Evolution,* 18: 65–66.

Mallet, J. and Willmott, K. (2003) Taxonomy: Renaissance or Tower of Babel? *Trends in Ecology and Evolution,* 10: 57–59.

Martinson, B.C., Anderson, M.S. and de Vries, R. (2005) Scientists behaving badly. *Nature,* 435: 737–738.

May, R. (2004) Tomorrow's taxonomy: Collecting new species in the field will remain the rate-limiting step. *Philosophical Transactions of the Royal Society of London, Series B,* 359: 733–734.

Meigen, J.W. (1830) *Systematische Beschreibung der bekannten europäischen zweiflügeligen Insekten*, vol. 6, Schulze, Theil.

Moritz, T. (2002) Building the biodiversity commons. *D-Lib Magazine,* 8(6), http://www.dlib.org/dlib/june02/moritz/06moritz.html.

Mueller, L., Tanksley, S.D., Giovanonni, J.J., van Eck, J., Stack, S., Choi, D., Kim, B.D., Chen, M., Cheng, Z., Li, C., Long, H., Xue, X., Seymour, G., Bishop, G., Bryan, G., Sharma, R., Khurana, J., Tyagi, A., Chattopadhyay, D., Singh, N.K., Stiekema, W., Londhout, P., Jesse, T., Lankhorst, R.K., Bouzayen, M., Shibata, T., Tabata, S., Granelli, A., Botella, M., Guiliano, G., Frusciante, L., Causse, M. and Zamir, D. (2005) The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comparative and Functional Genomics,* 6: 153–158.

Nic Lughadha, E. (2004) Towards a working list of all known plant species. *Philosophical Transactions of the Royal Society, Series B,* 359: 681–687.

Peralta, I.E., Knapp, S. and Spooner, D.M. (2008) Taxonomy of wild tomatoes and their relatives (*Solanum* sections *Lycopersicoides*, *Juglandifolia*, *Lycopersicon*; Solanaceae). *Systematic Botany Monographs,* in press.

Polaszek, A. et al. (2005) A universal register for animal names. *Nature,* 437: 477.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S. Kamoun, S., Sumlin, W.D. and Vogler, A.P. (2006) Evolutionary species delineation for the DNA taxonomy of undescribed insects. *Systematic Biology,* 55: 595–609.

Raven P.R. (2004) Taxonomy: Where are we now? *Philosophical Transactions of the Royal Society of London, Series B,* 359: 729–730.

Ride W.D.L. et al. (2000) *International Code of Zoological Nomenclature,* 4th ed., ITZN, London (also available at http://www.iczn.org/iczn/index.jsp).

Rubinoff, D., Sameron, S. and Will, K. (2005) Are plant barcodes a search for the Holy Grail? *Trends in Ecology and Evolution,* 21: 1–2.

Savolainen V. et al., eds. (2005) DNA barcoding of life: A theme issue. *Philosophical Transactions of the Royal Society of London, Series B*, 360: 1805–1975.

Scoble, M.J. (2003) Changing roles and perceptions in European natural history collections: From idiosyncracy to infrastructure. In *ENHSIN: The European Natural History Specimen Information Network* (ed M.J. Scoble). Published online at http://www.nhm.ac.uk/research-curation/projects/ENHSIN/documents/assets/ch1.pdf.

Scoble, M.J. (2004) Unitary or unified taxonomy? *Philosophical Transactions of the Royal Society of London, Series B,* 359: 699–710.

Seberg, O., Humphries, C.J., Knapp, S., Stevenson, D.W., Petersen, G., Scharff, N. and Andersen, N.M. (2003) Shortcuts in systematics? A comment on DNA taxonomy. *Trends in Ecology and Evolution,* 18, 63–65.

Seberg, O. and Petersen, G. Assembling the tree of life–magnitude, shortcuts, and pitfalls. In *Towards the Tree of Life: Taxonomy and Systematics of Large and Species-Rich Taxa* (eds T.R. Hodkinson, J.A.N. Parnell and S. Waldren), CRC Press, London, in press.

Service, R.F. (2006) The race for the $1000 genome. *Science,* 311: 1544–1546.

Smith G. (2004) The African Plants Initiative: A big step for continental taxonomy. *Taxon,* 53: 1023–1025.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2002) DNA points the way ahead for taxonomy. *Nature,* 418: 479.

Tautz, D. et al. (2003) A plea for DNA taxonomy. *Trends in Ecology and Evolution,* 18: 70–74.

Törnqvist T.E. (2002) Impact factors aren't top journals' sole attraction. *Nature,* 423: 480.

Turrill, W.B. and Milne-Redhead, E. (1952) Foreward and preface. In *Flora of Tropical East Africa* (eds W.B. Turrill and E. Milne-Redhead), Crown Agents for the Colonies, London, 54 pp.

Wheeler Q.D. (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society, Series B,* 359: 571–583.

Yoh, J. (2005) Brief introduction to the CDF experiment. CDF website at http://www-cdf.fnal.gov/events/cdfintro.html.

4 Planetary Biodiversity Inventories as Models for the New Taxonomy

Lawrence M. Page

CONTENTS

INTRODUCTION

The world faces a taxonomic crisis caused by the juxtaposition of the biodiversity crisis (National Science Board, 1989) and severe impediments to taxonomic research (Page et al., 2005). In 2003, the US National Science Foundation reacted to the taxonomic crisis by launching ‘planetary biodiversity inventories’, a new initiative intended to remove or reduce the impediments to taxonomic research. These impediments are insufficient taxonomic expertise, inadequate funding for research and isolation of resources required for completion of taxonomic research.

Planetary biodiversity inventories (PBIs) are global inventories of large clades of organisms predicted to contain many undescribed species or otherwise requiring major revision to complete their taxonomy. To accomplish the huge task of a global inventory of a large clade, each PBI must engage a multinational team of taxonomic experts and institutions with biological research collections.

The first competition for PBI funding was held in 2003. Four awards were made: one for the study of plant bugs (Miridae); one on slime molds (Eumycetozoa); one on the genus *Solanum*, a large plant genus containing nightshades, tomatoes and related plants; and one on catfishes (Siluriformes). Following is a discussion of the catfish project, which serves to illustrate the goals, organization and results of a PBI, and suggests that PBIs can serve as models for the New Taxonomy.

The principal goal of the All Catfish Species Inventory (ACSI) is to complete the taxonomy of Siluriformes, a monophyletic order of bony fishes. Completing the taxonomy includes describing undescribed taxa; completing generic and familial revisions on poorly known groups; developing identification keys, regional checklists and field guides; and making all taxonomically relevant information readily available through publications and websites. Principal investigators on the project are Jonathan W. Armbruster at Auburn University, John P. Friel at Cornell University, John G. Lundberg and Mark H. Sabaj at the Academy of Natural Sciences of Philadelphia, and Carl J. Ferraris, Jr. and Larry M. Page at the University of Florida Museum of Natural History.

Catfishes were chosen as the subject for a PBI because they are a monophyletic group (Fink and Fink, 1996; De Pinna, 1998; Saitoh et al., 2003), diverse (Burgess, 1989; Arratia et al., 2003) and worldwide in distribution. The group is predicted to include a large number of undescribed species, including those recognized but not scientifically described as well as species yet to be discovered. Also important in their selection for a PBI, catfishes were already under study by a large number of taxonomists who provide the nucleus of expertise necessary to identify specimens and revise higher level taxa. At the initiation of the project in 2003, 215 taxonomists and students signed on as participants; by the end of the project, the number of participants signed onto the project is expected to be about 400. Most participants are from North America and South America (Figure 4.1), two regions where systematic ichthyology is an active area of scientific research.

At the start of the project, 2855 named species of catfishes were considered to be valid. Although this number of species is small relative to that in some groups of invertebrates, catfishes constitute one of the largest orders of vertebrates. One in four

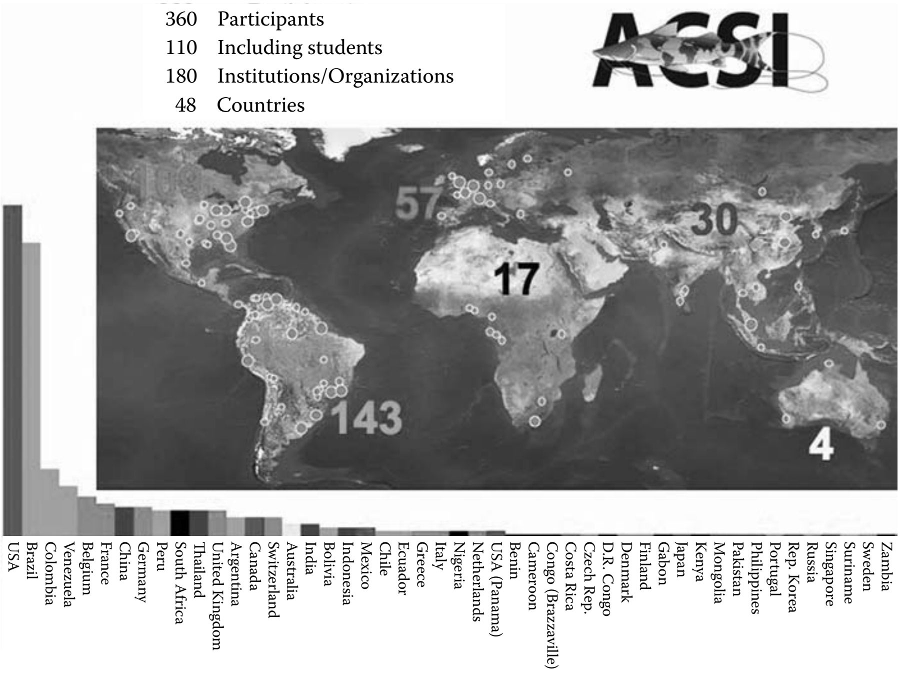


FIGURE 4.1 Distribution of scientists and students who have participated in the All Catfish Species Inventory in 2003–2005. Some dots represent localities with multiple participants and institutions.

species of all freshwater fishes is a catfish, one in 10 species of all fishes–marine, estuarine and freshwater–is a catfish, and one in 20 species of all species of vertebrates is a catfish. Even with this large number of described species, participants in 2003 estimated that 873–1750 species of catfishes remain to be described. These new species are represented by specimens of recognized but still unnamed species in institutional collections or are predicted to be discovered through additional fieldwork. The total number of species of catfishes to be recognized by the end of the project is predicted to be between 3600 and 4500.

The ACSI has a 5-year budget of $4.7 million to support taxonomic research on catfishes. Included are funds for fieldwork in poorly sampled regions likely to yield new species (primarily in tropical Africa, Asia and South America); visits to museums and other institutions with biological collections or taxonomic research programmes; assistance with illustrations, data analysis, and other tasks necessary to complete descriptions and revisions; and costs associated with publication (page charges, reprints and e-prints).

Three workshops, one each in Africa, Asia and South America, were held during the first year of the project to increase awareness and initiate collaborations. Fifty participants attended the workshop in Manaus, Brazil (held during the 2003 annual meeting of the American Society of Ichthyologists and Herpetologists), 15 attended the workshop in Singapore and 18 participants attended the workshop in Grahamstown, South Africa. Each workshop consisted of presentations on ACSI methods and support mechanisms, ongoing and planned research projects and roundtable discussions to describe and coordinate research.

Participants request support from ACSI by submitting short proposals following instructions on the ACSI website. About 90 per cent of the proposals submitted have been funded. ACSI has provided support thus far (end of the third year) for 118 projects led by participants other than the PIs. Most of the funding has gone to participants in South America (Table 4.1), which has the highest species-level diversity of catfishes and an active community of systematic ichthyologists. Most awards have supported fieldwork (38% of the awards) and visits to institutional collections (32%), or covered costs of publication (25%). Awards have ranged from USA $59 to $23,000 and have averaged $2959.

RESULTS

Increasing Funding for Research

To measure the effectiveness of increasing funding for taxonomic research, the number of new species of catfish described per year (Figure 4.2) was examined for the past 10 years. The number increased substantially in 2004 and 2005, suggesting that, although ACSI is only in its third year, providing even small increases in funding to researchers who usually lack funds can significantly accelerate taxonomic research.

Increasing Taxonomic Expertise

Alleviating the second major impediment to taxonomic research, insufficient taxonomic expertise, is a long-term goal. However, a large number of students (*c.*65 graduate

TABLE 4.1
Numbers of Proposals Funded by ACSI, by Country

		No. of Funded Proposals			
Continent or Region	**Country**	**Fieldwork, Visits to Institutions, etc.**	**Publication Support**	**Total by Country**	**Total by Continent**
South America					43
	Brazil	23	8	31	
	Colombia	8		8	
	Venezuela	3		3	
	Bolivia	1		1	
North America					43
	USA	14	27	41	
	Canada	1		1	
	Mexico		1	1	
Asia					7
	China	4		4	
	India	2		2	
	Indonesia	1		1	
Africa					5
	South Africa	3		3	
	Cameroon	1		1	
	Uganda	1		1	
Europe					5
	Belgium	2		2	
	France	2		2	
	Denmark	1		1	
Australia					2
	Australia	2		2	
South Pacific					1
	Fiji	1		1	
Total		70	36	106	106

Notes: Country refers to the location of the home institution of the lead participant on the proposal. Fieldwork, visits to institutions, etc. include assistance with other tasks necessary to complete description and revisions.

students) are working with ACSI PIs and participants and are being trained as the next generation of fish taxonomists. In this effort, PBIs are supplementing another successful NSF programme, Partnerships for Enhancing Expertise in Taxonomy (PEET), which prepares future generations of taxonomists by supporting taxonomists who train students to conduct monographic research (Rodman and Cody, 2003).

Graduate and undergraduate students involved in ACSI receive instruction in taxonomy, phylogenetics, biogeography, and natural history. Postdoctoral associates supported by ACSI participate heavily in research and work with foreign participants. All postdoctoral associates and graduate students participate in fieldwork and become familiar with the natural history and ecology of aquatic organisms.

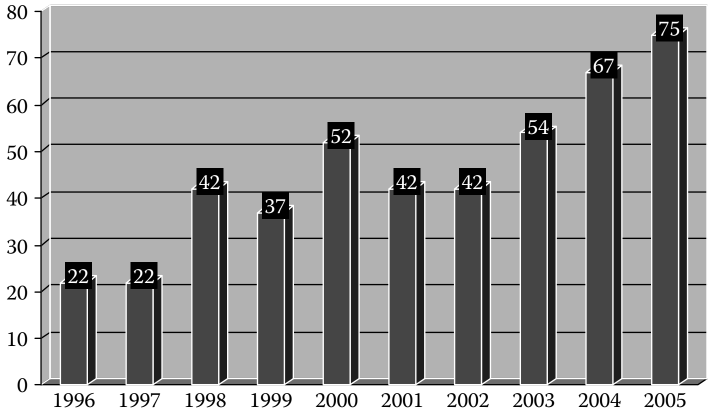


FIGURE 4.2 Number of new species of catfishes described per year. The All Catfish Species Inventory began in 2003.

They also participate in museum curation, species descriptions and other systematic research and manuscript preparation.

Removing the Isolation and Fragmentation of Resources

The third major impediment to taxonomic research, isolation and fragmentation of resources, is a long-standing problem that now can be largely overcome with the nearly universal access that scientists have to the Internet. ACSI has a website (http://silurus.acnatsci.org/) that provides a large amount of information that otherwise would not be available to most researchers and, in many cases, is directly responsible for the increase in completed species descriptions and revisions of catfish genera and families.

The ACSI website (Box 4.1) describes the objectives of ACSI, describes and illustrates catfish diversity and provides contact information for all participants. It also provides details about the scope and nature of participants' research projects, including taxa under study, so that collaborations may be initiated and data may be shared. Also posted on the website are a bibliography of all papers on the systematics of catfishes (*c*.3500 papers) and electronic copies of many old and otherwise difficult to obtain articles. These resources provide researchers with instant access to references they might not know about and to papers that otherwise might be unobtainable.

Box 4.1: Resources Available on the All Catfish Species Inventory Website, http://silurus.acnatsci.org/

An overview of the mission and goals of the All Catfish Species Inventory:

Information on catfish diversity, including lists of families (*N* = 35) and genera (*N* = 437) with numbers of species and images;

Details about the scope and nature of participants' research projects, including taxa under study and contact information;

Bibliography of all papers on the systematics of catfishes (*c.*3500 papers);
Electronic copies of old and difficult to obtain articles;
Digital images of catfishes, including primary types, live and freshly captured specimens;
An atlas of catfish morphology;
A list of repositories for name-bearing types of catfishes;
A list of acronyms for institutions with collections of catfishes;
Instructions for submitting research proposals to ACSI;
News and announcements relative to catfish taxonomy; and
Links to other websites with information relative to studies of catfishes, including the 'Catalog of Fishes', MorphoBank, Neodat II, DeepFin, Tree of Life, and sites supported by scientific societies, anglers and fish hobbyists.

Another resource available on the website is digital images of catfishes. Images of primary types of catfishes in all major institutional collections (about two-thirds of all primary types) are available. The remaining one-third are scattered among many small institutions. Most will be photographed by the end of 2007. These images at least enable a researcher to identify the institutions that are critical to visit and at best allow a researcher to complete a study without traveling long distances to examine specimens. Images of live and freshly captured specimens and of unusual species also are available to facilitate descriptions, as is an online atlas of catfish morphology.

To facilitate communication among ACSI participants, an electronic mail list server, Siluri-Net, has been established at the Academy of Natural Sciences of Philadelphia. Communications through Siluri-Net function as a project newsletter, allowing ACSI participants and other qualified professionals and students to describe research projects, query participants about the availability of specimens or literature, announce opportunities for scientists to work together on fieldwork and species descriptions, and discuss educational and conservation topics related to catfishes.

In addition to improving access to taxonomic information and communication, ACSI makes it easier for researchers to publish results of their studies. Rapid publication of species descriptions historically has been hampered by the limited availability of publication outlets, the slow rate of processing manuscripts and high cost. This is true especially for taxonomists working in less developed countries. In addition to providing funds to cover costs of publication, ACSI provides editorial assistance to participants who choose to publish in *Zootaxa*, an international electronic journal for taxonomic studies (www.mapress.com/zootaxa). The number of papers on catfishes published in *Zootaxa* has increased dramatically following the initiation of editorial and financial assistance from ACSI (Figure 4.3).

ACSI has two more years to meet its original objectives; however, early signs are that it is a success. More species are being described, and taxonomic information previously unavailable is being published and distributed electronically. Major revisions have appeared (Vari et al., 2005; Sabaj, 2005), exciting discoveries have been announced–including the discovery of a new family of catfishes (Rodiles-Hernández et al., 2005)–and phylogenetic and diagnostic information for large clades is being published (Armbruster, 2004; Thomson and Page, 2006).

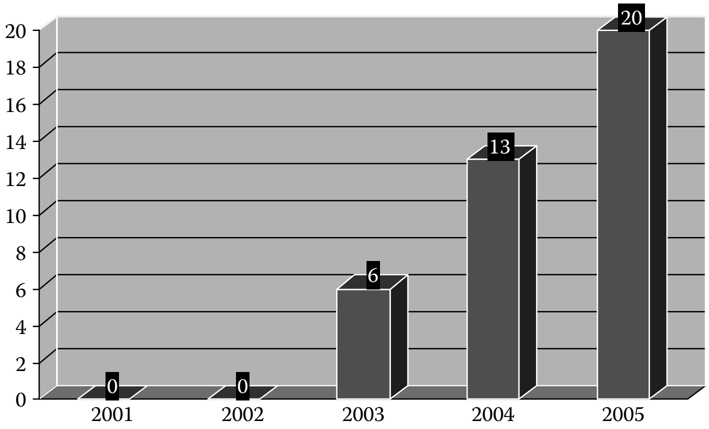


FIGURE 4.3 Numbers of catfish papers published in *Zootaxa* per year. *Zootaxa* was launched in 2001 and the All Catfish Species Inventory began in 2003.

Given the severity of the taxonomic crisis and the success of the catfish PBI, it is evident that more PBIs should be funded by the US National Science Foundation and that PBI-like initiatives should be launched elsewhere in the world. A short time remains to document the biological diversity of our planet, and biological surveys and inventories must be a high priority for science. PBIs are a means to address that priority.

LINNE, the topic of the next chapter, is an initiative to accelerate taxonomic research and improve biological collection infrastructure so that reliable information on biological diversity is available to all branches of science and society. LINNE promises to reduce impediments to taxonomy even more effectively than do PBIs.

ACKNOWLEDGEMENTS

The All Catfish Species Inventory is funded by US National Science Foundation award DEB-0315963. Griffin Sheehy compiled the information for Table 4.1 and Figure 4.1. John Lundberg provided the map for Figure 4.1. Carl Ferraris provided the data for Figure 4.2.

REFERENCES

Armbruster, J.W. (2004) Phylogenetic relationships of the suckermouth armoured catfishes (Loricariidae) with emphasis on the Hypostominae and the Ancistrinae. *Zoological Journal of the Linnean Society,* 141: 1–80.

Arratia, G., Kapoor, B.G., Chardon, M. and Diogo, R. (2003) *Catfishes,* Science Publishers, Inc., Endfield, NH, Vols. 1 and 2, 812 pp.

Burgess, W.E. (1989) *An Atlas of Freshwater and Marine Catfishes,* T.F.H. Publications, Neptune City, NJ, 784 pp.

De Pinna, M.C.C. (1998) Phylogenetic relationships of neotropical Siluriformes (Teleostei: Ostariophysi): Historical overview and synthesis of hypotheses. In *Phylogeny and Classification of Neotropical Fishes* (eds L.R. Malabarba, R.E. Reis, R.P. Vari, Z.M.S. Lucena and C.A.S. Lucena), EDIPUCRS, Porto Alegre, Brazil, pp. 279–330.

Fink, S.V. and Fink, W.L. (1996) Interrelationships of ostariophysans fishes (Teleostei). In *Interrelationships of Fishes* (eds M.L.J. Stiassny, L.R. Parenti and G.D. Johnson), Academic Press, New York, pp. 209–247.

National Science Board (1989) Loss of biological diversity: A global crisis requiring international solutions. Report NSB 89-171. National Science Foundation, Washington, D.C.

Page, L.M., Bart, H. L., Jr., Beaman, R., Bohs, L., Deck, L.T., Funk, V.A., Lipscomb, D., Mares, M.A., Prather, L.A., Stevenson, J., Wheeler, Q.D., Woolley, J.B. and Stevenson, D.W. (2005) *LINNE: Legacy Infrastructure Network for Natural Environments,* Illinois Natural History Survey, Champaign, IL., 16 pp.

Rodiles-Hernández, R., Hendrickson, D.A., Lundberg, J.G. and Humphries, J.M. (2005) *Lacantunia enigmatica* (Teleostei: Siluriformes), a new and phylogenetically puzzling freshwater fish from Mesoamerica. *Zootaxa,* 1000: 1–24.

Rodman, J.E. and Cody, J.H. (2003) The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. *Systematic Biology,* 52: 428–435.

Sabaj, M.H. Taxonomic assessment of *Leptodoras* (Siluriformes: Doradidae) with descriptions of three new species. *Neotropical Ichthyology,* 3: 637–678.

Saitoh, K., Miya, M., Inoue, J.G., Ishiguro, N.B. and Nishida, M. (2003) Mitochondrial genomics of ostariophysan fishes: Perspectives on phylogeny and biogeography. *Journal of Molecular Evolution,* 56: 464–472.

Thomson, A.W. and Page, L.M. (2006) Genera of the Asian catfish families Sisoridae and Erethistidae. *Zootaxa,* 1345: 1–96.

Vari, R.P., Ferraris, C.J., Jr. and de Pinna, M.C.C. (2005) The neotropical whale catfishes (Siluriformes: Cetopsidae: Cetopsinae), a revisionary study. *Neotropical Ichthyology,* 3: 127–238.

5 On the Use of Taxonomic Concepts in Support of Biodiversity Research and Taxonomy

Nico M. Franz, Robert K. Peet and Alan S. Weakley

CONTENTS

> Linnaean nomenclature is stable enough to say what we know, flexible enough to accommodate what we learn; independent of specific theory, yet reflective of known empirical data; compatible with phylogenetic theory, but not a slave to it; particular enough for precise communication, general enough to reflect refuted hypotheses.
>
> **(Wheeler, 2004, p. 577)**

INTRODUCTION

The current system of nomenclature works well enough for many users and purposes. Linnaean names are both responsive to certain changes in taxonomic perspective and fairly stable. The former is necessary so that taxonomists can express what they learn about nature's entities and their relationships. The latter helps users such as ecologists understand each other's results–even if they are separated in space and time. Linnaean names have successfully played the role of a working compromise for 250 years.

New developments are beginning to challenge the view that the Linnaean system of nomenclature is able to satisfy the requirements of the scientific community. Future biodiversity research will become increasingly dependent upon distributed data networks, scientific workflows and ontology-driven mechanisms for resolving a broad spectrum of primary data (Ludäscher et al., 2005, 2006). Biodiversity informatics must therefore provide an information technology infrastructure to support such complex tasks (Page, 2005; Michener et al., 2007).

A prime use case for developers of biodiversity informatics technology is the ecological niche modelling (predicting of geographic ranges–past, present and future) of a specific set of taxa based on museum specimen data (Soberón and Peterson, 2004). Taxonomic resolution is an important part of this use case, yet Linnaean names by themselves are often not precise enough to resolve data to the level required. In what context these issues occur, why they exist, how significant they are, and what ideas and tools are being developed to solve them is the subject of this paper. Throughout, the 'taxonomic concept' approach is presented as a solution not only to problems in biodiversity research, but also for the long-term management of evolving perspectives in taxonomy proper.

THE CHALLENGE OF TAXONOMIC RESOLUTION IN A COMPLEX BIODIVERSITY ANALYSIS

Taxonomic resolution presents a significant challenge in a wide range of biodiversity studies. Consider, for example, the task of predicting the distributions of species of North American mammals using a workflow analysis. Two major sources of input are needed to run such an analysis. One is a list of individual specimens as recorded by museum databases and made accessible, for example, via the *Mammal Networked Information System.*[1] A user of the workflow infrastructure may thus call up approximately 1.5 million records, an estimated 10–20 per cent of which have latitudinal/longitudinal data in decimal format. A typical record for the striped skunk would read '*Mephitis mephitis;* 42.456°N; –84.013°W'. The other input source is a set of georeferenced environmental variables such as topographic indices, historical climate measurements (precipitation, cloud cover, temperature, etc.) and vegetation type information. The entirety of these variables makes up the ecological niche that an individual taxon can presumably inhabit. Future distributions are then modelled using a generic algorithm for rule-set prediction under varying global climate scenarios (Peterson et al., 2002). The output is a color map with range predictions.

Like most biodiversity studies, the aforementioned mammal workflow has a critical taxonomic component. Suppose a researcher wants to predict the distributions for two different species of skunk. The process of importing the museum records must therefore produce all relevant distribution data for two separate biological entities– and nothing else. If the query fails to retrieve all data, the analysis loses power. If irrelevant records are included or the delimitations of taxa are blurred, then the results might be false. Reliable niche predictions require precise taxonomic resolution.

When the researcher enters the names '*Mephitis mephitis*' and '*Spilogale purtorius*' to assemble all records for two kinds of skunk, he or she has to make several assumptions. The museum records might cover the entire North American region. Many date back to the nineteenth century. One assumption is that the specimens were identified correctly according to the then preferred reference works. Although the quality of identification can vary with the taxa under study (Meier and Dikow, 2004), this is not something one can rectify easily from a remote location.

Even if the identifications were carried out properly, a number of questions remain. For instance, is it safe to assume that data linked to different names belong to separate taxa? And, vice versa, is it safe to assume that data linked to the same name may be pooled into one list? Furthermore, is everything that used to be labelled '*Mephitis*' still part of *Mephitis* as recognized now? Will a query for '*Mephitis*' necessarily yield all records pertinent to the analysis? In each case the answer is likely negative, so the researcher will have to take additional steps to resolve the names to meaningful biological entities. This task may include recognizing and correcting for variant name spellings (Chapman, 2005), adding records with names that are subordinate in the Linnaean hierarchy and–most importantly–identifying and merging records labelled with synonyms. Although these resolution steps will greatly improve the analysis, two significant problems remain. First, any decision to rectify, separate or merge data will be made in accordance with (at least) one authoritative taxonomic treatment. The latter may play the role of a 'standard' now, but will be outdated in a few years. The possibility to interpret and reutilize the data in the future will therefore decrease (see also Michener et al., 1997). Second, for reasons that will be explained hereafter, the practice of merging or disjoining data on the basis of synonymy is inherently too imprecise to meet all resolution needs. In short, the conventional approach to taxonomic resolution via Linnaean names, hierarchy and synonymy relationships is not an optimal long-term solution.

THE RELATIONSHIP OF LINNAEAN NAMES AND EVOLVING TAXONOMIC PERSPECTIVES

Today's nomenclatural practice relies on methods such as the designation of type specimens and the principle of priority. Although sometimes under attack, these conventions have a long record of improving communication about nature. They are open to more than one theoretical interpretation (Farber, 1976; Stevens, 1984), thereby contributing to the transgenerational character of the Linnaean system. Nevertheless, because the rules of nomenclature were designed to strike a working balance, continuity and change in naming are not inextricably linked to the evolution of taxonomic perspectives. Not every new taxonomic judgement is labelled with a unique name, and not every name change reflects a revised view of taxonomic circumscription or relationship. This insight is old and might seem trivial, since all humans are accustomed to updating terms or revising their meanings from time to time. However, in the context of achieving precise taxonomic resolution, it is appropriate to examine the connection of nomenclature and taxonomy more closely.

Taxonomists and most other biologists are familiar with the particularities of naming versus delimiting taxa. For example, the senior author recently published an analysis of the weevil tribe Derelomini Lacordaire (Franz, 2006). The tribe now includes 11 genera that were placed elsewhere in the preceding weevil *Catalogue* (Alonso-Zarazaga and Lyal, 1999). It also excludes six genera that used to be part of the tribe. Only 2 of the 41 currently recognized genera, one of them now under a different name, were assigned to the Derelomini (then spelled *Dérélomides*) when the name was first defined in the mid-nineteenth century. Future taxonomic updates such as revised diagnoses, additions and subtractions of non-type elements will change the meaning of 'Derelomini', but not the name itself. In such cases the name and its meaning evolve independently.

The partial disconnect of nomenclature and taxonomy may be illustrated with a contrived example (Figure 5.1; see also Kennedy et al., 2005; Page, 2006). Suppose that in 1798 Fabricius named a new genus *Fantasia* F. and species *F. prima* F., based on a heterogeneous series of specimens.[2] One specimen was designated as the holotype. In 1903 Champion decided that parts of the series belong to two additional distinct species, named *F. secunda* Champion and *F. tertia* Champion. Two more holotypes were selected to represent the new entities. In 1948 Bondar reassigned the specimens 'unevenly' to two of the three existing names. A heterotypic synonym *F. secunda* was created for *F. prima,* which has priority. Also, a subset of the *F. secunda* specimens (according to Champion) was renamed *F. tertia.* Finally, in 2000 Afterall analysed parts of the original 1798 material, as well as newly collected specimens with somewhat deviating features. The specimen circumscription of *F. prima* is now more inclusive in comparison to 1798 or 1903, and overlapping with 1948. The name *F. secunda* is resurrected to apply to Champion's holotype and several other specimens. The material named *F. tertia* by Champion is judged

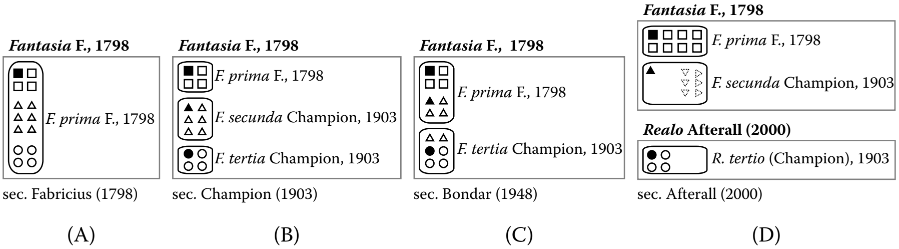


FIGURE 5.1 Sequence of four treatments of the hypothetical taxon *Fantasia* F., authored by (A) Fabricius (1798); (B) Champion (1903); (C) Bondar (1948); and (D) Afterall (2000). Individual specimens are represented with the symbols □, △, ○, etc. The relevant nomenclatural types for species and higher level taxa are shown as ■, ▲ and ●. See text for further details.

sufficiently distinctive to merit the creation of a new genus name *Realo* Afterall. The epithet for its type species is changed to *R. tertio* (Champion) to match the new gender.

The example clarifies the effect of the method of types and nomenclatural priority. For instance, there are at least three different perspectives on what the name *F. prima* means. They share the same holotype, yet the non-type elements can vary greatly in their circumscription. On the other hand, Champion's *F. tertia* and Afterall's *R. tertio* are different names with the same meaning. But this does not mean that synonymy, which is essentially a two-point comparison, can always provide the required level of taxonomic resolution. The relationships of names and meanings become still more difficult to trace if strictly nomenclatural errors are considered (spelling, availability, validity, etc.).

With the important exceptions of the genus/species link and ranks, Linnaean names change in response to new taxonomic judgements only to the extent that the uniqueness and priority of primary types is affected. Whatever 'surrounds' these types and otherwise lacks priority may undergo rearrangement without triggering additional naming acts. And therein lies the inherent imprecision of Linnaean names. A researcher aiming for accurate niche modelling results must understand which circumscription of *F. prima* was used to label the museum records of interest. Was it that of 1798, 1903 or 1948? The identification label '*F. prima* F.' is likely not enough to retrieve a taxonomically congruent set of records. Reliable inferences of future distributions will have to depend upon more precise semantics than those offered by Linnaean names and synonymy alone.

INTRODUCING TAXONOMIC CONCEPTS

The solution to the preceding challenge is to specify the author and publication where the meaning of *F. prima* was defined or redefined. This solution is called the 'taxonomic concept' approach. It is already implemented in select taxonomic databases.[3] A taxonomic concept is the underlying meaning, or referential extension, of a scientific name as stated by a particular author in a particular publication. It represents the author's full-blown view of how the name is supposed to reach out to objects in nature.[4] It is a direct reflection of what has been written, illustrated and deposited by a taxonomist, regardless of his or her theoretical orientation.

In order to label the different usages of a name, Berendsohn (1995) proposed the term 'sec.' from the Latin *secundum,* or 'according to'. The 'sec.' is preceded by the full Linnaean name and followed by the specific author and publication. Two examples are *F. prima* F. sec. Fabricius (1798; the original concept) and *F. prima* F. sec. Afterall (2000; the most recent concept). Thus, the concept approach allows one to address the various published meanings of the name *F. prima* F. It is now possible to trace their evolution through time.

AN EMERGING LANGUAGE FOR CONCEPT RELATIONSHIPS

As soon as the multiple usages of a name are assigned to their source, each of them may be reconnected in ways that are more precise than type-based definitions and

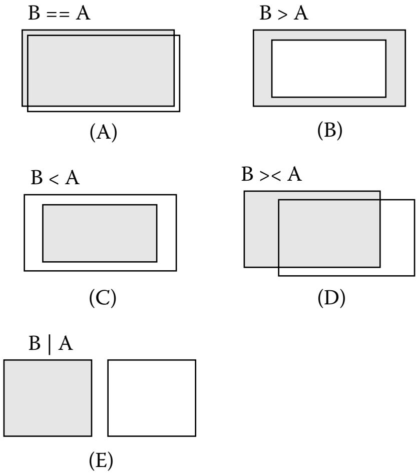


FIGURE 5.2 Schematic representation of the five basic kinds of concept relationships. The referential extension of concept A is indicated by the white rectangle, whereas that of concept B is indicated by the shaded rectangle. (A) Congruence; (B) B is more inclusive than A; (C) B is less inclusive than A; (D) B overlaps with A; and (E) B excludes A.

synonymy relationships.[5] Five basic symbols and meanings derived from set theory are used for comparing two concepts A and B (Figure 5.2): B is congruent with A, B is more inclusive than A, B is less inclusive than A, B overlaps with A and B excludes A. The meanings should be viewed as mutually exclusive in order to maximize their usefulness (Koperski et al., 2000). Hence, 'overlap' means that each concept has some unique (non-shared) elements in addition to shared ones. A relationship assessment may take everything into consideration that is tied to the respective concepts, including sets of specimens, subordinate concepts and character circumscriptions. Explanatory comments can complement the assessments, especially in the case of incongruence.

Several additional terms have proven useful for expressing concept relationships. Their meanings and applications are summarized in Table 5.1. Most high-quality concepts will have both a diagnosis (intensional component) and a list of included subelements (ostensive component). These two aspects tend to complement each other, although the message they send need not be the same. Diagnoses reach out to as of yet unexamined objects; specimens are sometimes mislabelled, etc. Assessing concept relationships is a non-trivial task left for taxonomic experts.

Returning to the preceding case (Figure 5.1), one can now specify the taxonomic changes within *Fantasia* F. using the concept approach. For instance, *F. prima* F. sec. Fabricius (1798) is more inclusive (>) than *F. prima* F. sec. Champion (1903). Champion's other two concepts must be added to obtain congruence: *F. prima* F. sec. Fabricius (1798) is congruent (==) with the sum of *F. prima* F. sec. Champion (1903) plus (+) *F. secunda* Champion sec. Champion (1903) plus (+) *F. tertia* Champion sec. Champion (1903). In another comparison, *F. secunda* Champion sec. Champion (1903) overlaps (><) with *F. tertia* Champion sec. Bondar (1948). The two concepts share some non-type elements. Finally, *F. prima* F. sec. Champion (1903) is

TABLE 5.1
Additional Terms to Express Concept Relationships

Symbol or Term	Meaning	Example
Is parent of	A concept is superordinate to another within the same hierarchy.	A is a parent of B
Is child of	A concept is subordinate to another within the same hierarchy.	C is a child of D
+ (Plus)	The extensions of two concepts are added together.	A + B == C
– (Minus)	The extension of a concept is subtracted from another.	B == C – D
AND	Permits the concatenation of multiple valid assertions.	A == (INT) B AND A > (OST) B
OR	Permits the expression of uncertainty via alternative assertions.	A == B OR A > B
INT (intensional)	The relationship is based only on diagnostic properties.	A == (INT) B
OST (ostensive)	The relationship is based only on constituent subelements.	A > (OST) B

Note: See also Figure 5.2.

intensionally congruent (== INT) with *F. prima* F. sec. Afterall (2000), and also *F. prima* F. sec. Champion (1903) is less inclusive ostensively (< OST) than *F. prima* F. sec. Afterall (2000). The latter author listed more elements, albeit of the same kind as Champion's. The intensional/ostensive distinction is useful in particular at higher taxonomic levels.

LONG-TERM TAXONOMIC RESOLUTION USING THE CONCEPT APPROACH

The imperfect connection of nomenclatural and taxonomic adjustments over time mandates that long-term taxonomic resolution for biodiversity research be based not just on type-driven name definitions, but also on the more powerful concept relationships. The vision for implementing such a service is as follows. The future storage and integration of ecological data will be made possible via a comprehensive metadata approach (Michener and Brunt, 2000). An integral part of this approach is the linking of primary observations to taxonomic concepts. This means that biodiversity researchers, when submitting their data to a networked database, will be required to *identify* these observations to sets of well specified concepts. As an example, the conventional entry '*Mephitis mephitis* Schreber' would be submitted as '*Mephitis mephitis* Schreber sec. Wilson and Reeder (1993), if the latter were the reference consulted in the identification process. Researchers may equip their identifications with an assessment of certainty. Eventually, the authoritative concepts will need to receive unique identifiers, such as those of the *Digital Object Identifier* system (Paskin, 2005).

In a separate process, taxonomic concepts must be related to each other using the preceding language for concept relationships (e.g. *Mephitis mephitis* sec. Wilson and Reader [1993] is more inclusive [>] than *Mephitis major* sec. Howell [1901]). The integration and dissociation of data are then based upon the relationships, with some flexibility to match the resolution needs of each analysis. The primary biodiversity data will remain resolvable for the long term, so long as the originally referenced concepts are well specified and connectable to elements in succeeding classifications.

Biodiversity studies that pay attention to the dynamics of taxonomy often yield astonishing results. For instance, Peterson and Navarro-Sigüenza (1999) analysed avian endemism in Mexico using two alternative taxonomies. Under the biological species concept, 101 endemic species were obtained, with most endemics concentrated in the southern and western montane areas. Application of the phylogenetic species concept, in turn, produced 249 endemic species, a majority of which occurs in the western lowlands and mountains. Selecting one classification over the other therefore greatly affects conservation priorities. The concept approach is well suited to expose such critical interdependencies. Analyses similar to those of Peterson and Navarro-Sigüenza (1999) present a powerful way to convince the ecological user community of its utility.

THE TAXONOMIC CONCEPT APPROACH PUT IN PRACTICE

In order to benefit biodiversity research, the concept approach must above all make practical sense for taxonomists. The implementation of concept taxonomy in two otherwise traditional treatments indicates that this is so. The particularities of each treatment will be reviewed briefly.

The Checklist of German Mosses (Koperski et al., 2000) is a pioneering effort in concept taxonomy. According to the authors' perspective, 1548 names and concepts are accepted at the generic and lower levels (see Geoffroy and Berendsohn, 2003). An additional 6996 invalid names (i.e. homotypic and heterotypic synonyms) are listed. The names and synonyms are derived from an analysis of 11 major taxonomic reference works for Central European mosses, the oldest dating back to 1927. The authors combine the 8544 names and 12 references for a total of 24,390 cited taxonomic concepts. They established 7891 concept relationships connecting each member in the accepted pool of concepts to one or more suitable predecessors. In short, the Checklist provides insight into the evolution of German moss classifications over a time span of 73 years.

The format adopted by Koperski et al. (2000) places conventional information about nomenclature alongside the new concept relationships (Figure 5.3A). For each entry of an accepted concept the authors provide the complete original citation. They also list the existing synonyms, either homotypic or heterotypic, as well as other invalid or unavailable names ('auct.'). The entry is then completed with a series of concept relationships (typically less than 10) connecting the accepted concept to its congruent or (partly) incongruent predecessors. Often notes are added to explain particular judgements and kinds of incongruence. At the end of a genus-level entry, all unaccepted names are assigned to their valid counterparts (Figure 5.3A). These assignments are necessary due to the fact that there may be many-to-many relationships between invalid and valid names. In summary, the German moss *Checklist*

	Dicranum fuscescens Sm.	[sec. Koperski et al., 2000]
	Flora Britannica, 1804: 27	[nomenclatural source]
	= *Dicranum congestum* Brid.	[heterotypic synonym]
==	*Dicranum fuscescens* Sm.	sec. Corley et al. (1981, 1991)
==	*Dicranum fuscescens* Sm. Ludwig et al. (see there) refer to the concept of Corley et al.	sec. Ludwig et al. (1996)
==	*Dicranum fuscescens* var. *eu-fuscescens* Mönk.	sec. Mönkemeyer (1927)
<	*Dicranum fuscescens* Turner Includes D. flexicaule (see comments there).	sec. Frahm and Frey (1992)
<	*Dicranum fuscescens* Turner Includes D. flexicaule (see comments there).	sec. Mönkemeyer (1927)
<	*Dicranum fuscescens* Sm. Includes D. flexicaule in the type variety (cf. morphological account).	sec. Smith (1980)
>	*Dicranum fuscescens* var. *congestum* (Brid.) Husn. This taxon is evidently a montane growth form of D. fuscescens.	sec. Smith (1980)
><	*Dicranum fuscescens* var. *fuscescens*	sec. Smith (1980)

Dicranum congestum Brid.	*Dicranum fuscescens* Sm.
Dicranum enerve Hedw.	*Paraleucobryum enerve* (Hedw.) Schimp.
Dicranum palustre Bruch & Schimp.	*Dicranum bonjeanii* De Not.
...	

(A)

Aureolaria flava (Linnaeus) Farwell var. **flava**, Estearn Smooth Oak-leach. Pd, Mt, Cp (GA, NC, SC, VA): oak forests and woodlands; common. August-September; September-October. ME west to MN, south to GA, FL, and AL. Var. reticulata (Rafinesque) Pennell, of the southeastern Coastal Plain, needs additional study. It is alleged to differ n its lower leaves entire, dentate, or divided less than 1/2 way to the midrib (vs. deeply pinnatifid-divided). [== C, G, K; < A. flava -- RAB, W; > Gerardia flava Linnaeus var. flava -- F; > A. flava ssp. typica -- P; >< flava ssp. flava -- S; > A. flava spp. reticulata (Rafinesque) Pennell -- P, S]

(B)

FIGURE 5.3 Exemplary representational conventions for implementing concept taxonomy in practice (slightly modified). (A) Entry for the concept of *Dicranum fuscescens* Sm. sec. Koperski et al. (2000), including eight (partially) annotated concept relationships and three exemplary assignments of invalid to valid names. (B) Entry for the concept of *Aureolaria flava* (Linnaeus) Farwell var. *flava* sec. Weakley (2006). Data on bionomics are followed by 10 concept relationships displayed in square brackets [ ]. 'C, G, K, RAB, W', etc. are abbreviations for preceding reference works, and '--' is used instead of 'sec.'

offers its users more nomenclatural and taxonomic information than any traditional work of this scope.

The Flora of the Carolinas project (Weakley, 2006) is another powerful example of concept taxonomy put in practice. This treatment considers approximately 6300 names and concepts as valid. The latter are connected to taxonomic elements of up to

10 earlier reference works published between 1933 and 2005. More than 40,000 concept relationships connect the accepted concepts to their predecessors (Figure 5.3B). The format for displaying the relationships dovetails neatly with the remaining content and greatly enhances the taxonomic value of this publication.

Several additional implementations of the concept approach are currently under way (see also the earlier footnote on select taxonomic databases). For instance, the major repository for prokaryote nomenclature and taxonomy is adopting concepts in combination with unique identifiers (Garrity and Lyons, 2003). North American vascular plant databases are also preparing for this transition. Smaller scale projects such as a concept-based database of angelfishes (R.L. Pyle, personal communication) are emerging at various locations. These efforts underscore the general practicality of the concept approach.

WHAT CONCEPT RELATIONSHIPS SAY ABOUT THE PRECISION AND RELIABILITY OF LINNAEAN NAMES

The applications of concept taxonomy offer new and quantitative insights into the performance of Linnaean names. Specifically, evaluations of the relative abundance of congruent versus incongruent relationships reflect on the precision and reliability of names over a given time span. Such assessments may be carried out as a series of two-point comparisons (i.e. reaching out repeatedly from a current set of concepts to multiple preceding sets) or through examination of entire 'concept lineages' in chronological order. The results are then contrasted with parallel analyses of stability and change in naming alone.

Geoffroy and Berendsohn (2003) analysed the moss data published by Koperski et al. (2000) along these lines. Taking the 1548 therein recognized concepts as the accepted standard, they calculated that 1509 concepts (97.5%) had at least one congruent predecessor. Many concepts had additional incongruent predecessors (Table 5.2). At a finer level of resolution, 550 concepts (35.5%) were likely taxonomically stable

TABLE 5.2
Distribution of Five Kinds of Relationship Linking the 1548 Accepted Concepts in Koperski et al. (2000) to Their Respective Predecessors[a]

Relationship	No. of Concepts	Per Cent of Concepts
==	1509	97.5
>	267	17.2
<	515	33.3
><	90	5.8
|	11	0.7

[a] See Geoffroy and Berendsohn (2003).

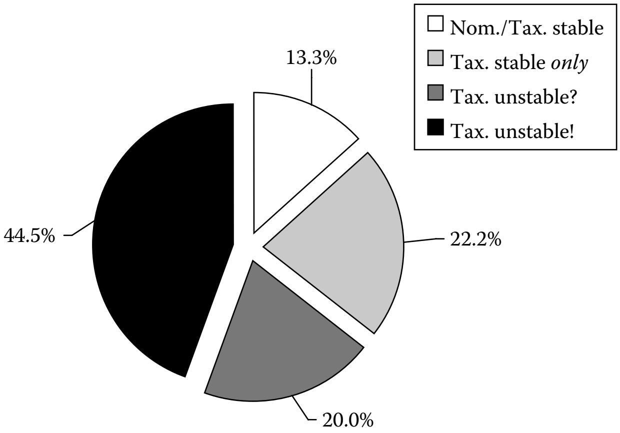


FIGURE 5.4 Pie diagram showing the percent distribution of nomenclaturally and/or taxonomically stable and unstable concepts analysed in Koperski et al.'s (2000) *Checklist* (N = 1548 accepted concepts). (Data from Geoffroy, M. and Berendsohn, W.G., 2003, *Schrifteneihe für Vegetationskunde,* 39: 5–14.)

from 1927 to 2000, citing only homotypic synonyms and congruent relationships to previously established concepts. As many as 310 concepts (20.0%) were potentially unstable due to heterotypic synonyms or misapplied names. And no less than 688 concepts (44.5%) were explicitly unstable, citing one or more incongruent relationships. Within the latter group of unstable concepts, 530 concepts (77.0%) referenced a single kind of incongruence, 122 concepts (17.7%) mentioned two kinds, 35 concepts (5.1%) cited three kinds (see also Figure 5.3A), and one concept (0.1%) had all four kinds. In what is perhaps the most telling statistic from this analysis, the authors concluded that only 207 concepts (13.3%) out of a total of 1548 concepts have remained the same in name *and* taxonomic meaning throughout the past 73 years (Figure 5.4). This value is low, especially if one considers how well this particular flora was studied by 1927. Biodiversity researchers who need to integrate data across the analysed time period may trust a name roughly one out of eight times.

Weakley (2006) carried out similar analyses with relationships originating from the Flora of the Carolinas project. The two-point concept comparisons between the Flora's perspective and eight relevant predecessors yielded 77–94 per cent congruence (Table 5.3). Not surprisingly, the percentage of incongruent concepts increases with time. The overwhelming majority of incongruent relationships were of the '>' or '<' kind. The author also provided data on stability in name *and* taxonomic meaning, which ranged from 55 to 88 per cent in the eight comparisons. These numbers seem more reassuring than the results for German mosses. Yet this impression will change when entire concept lineages are analysed. An example of concept evolution in *Andropogon* L. sec. Weakley (2006) shows how poorly the names and taxonomic perspectives match among succeeding treatments (Figure 5.5). Using the concept approach is required to discover such discrepancies in the first place and to properly realign them.

sec. Hackel (1889)	sec. Small (1933)	sec. Blomquist (1948)	sec. Hitchcock & C. (1950)	sec. RAD (1968)	sec. Godfrey and W. (1979)	sec. Campbell (1983)	sec. Weakley (2005)
A. virginicus var. *glaucus* subvar. *glaucus*	*A. capillipes*	*A. capillipes*	*A. capillipes*	*A. virginicus*	*A. capillipes*	*A. virginicus* var. *glaucus* "drylands variant"	*A. capillipes* var. *capillipes*
A. virginicus var. *glaucus* subvar. *dealbatus*	*A. capillipes*	*A. capillipes*	*A. capillipes*	*A. virginicus*	*A. capillipes*	*A. virginicus* var. *glaucus* "wetlands variant"	*A. capillipes* var. *dealbatus*
A. virginicus var. *viridis* subvar. genuinus	*A. virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus* var. *virginicus* "old-field variant"	*A. virginicus* var. *virginicus*
A. virginicus var. *viridis* subvar. *genuinus*	*A. virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus* var. *virginicus* "smooth variant"	*A. virginicus* var. *virginicus*
A. virginicus var. *viridis* subvar. *genuinus*	*A. virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus*	*A. virginicus* var. *virginicus*	*A. virginicus* var. *decipiens*	*A. virginicus* var. *decipiens*
A. macrourus var. *glaucopsis*	*A. glomeratus*	*A. virginicus* var. *glaucopsis*	*A. virginicus* var. *glaucopsis*	*A. virginicus*	*A. glaucopsis*	*A. glomeratus* var. *glaucopsis*	*A. glaucopsis*
A. macrourus var. *hirsutior*	*A. glomeratus*	?	*A. virginicus* var. *hirsutior*	*A. virginicus*	*A. virginicus* var. *abbreviatus*	*A. glomeratus* var. *hirsutior*	*A. glomeratus* var. *hirsutior*
A. macrourus var. *abbreviatus*	*A. glomeratus*	*A. glomeratus*	*A. glomeratus*	*A. virginicus*	*A. virginicus* var. *abbreviatus*	*A. glomeratus* var. glomeratus	*A. glomeratus* var. *glomeratus*
A. macrourus var. *genuinus*	*A. glomeratus*	*A. virginicus* var. *tenuispatheus*	*A. glomeratus*	*A. virginicus*	*A. virginicus* var. *abbreviatus*	*A. glomeratus* var. *pumilus*	*A. tenuispatheus*

FIGURE 5.5 Concept evolution in the grass genus *Andropogon* L. according to eight succeeding treatments. (Data from Weakley, A.S., Flora of the Carolinas, Virginia, and Georgia. Working draft of January 17, 2006. Available at http://www.herbarium.unc.edu/flora.htm 1015 pp.). Each column contains a coherent perspective, and each row represents a congruent concept–irrespective of the names used to label the individual cells. Taxonomic concepts whose circumscriptions are shared among multiple authors are colored with unique patterns of shading, whereas concepts unique to a single source are white.

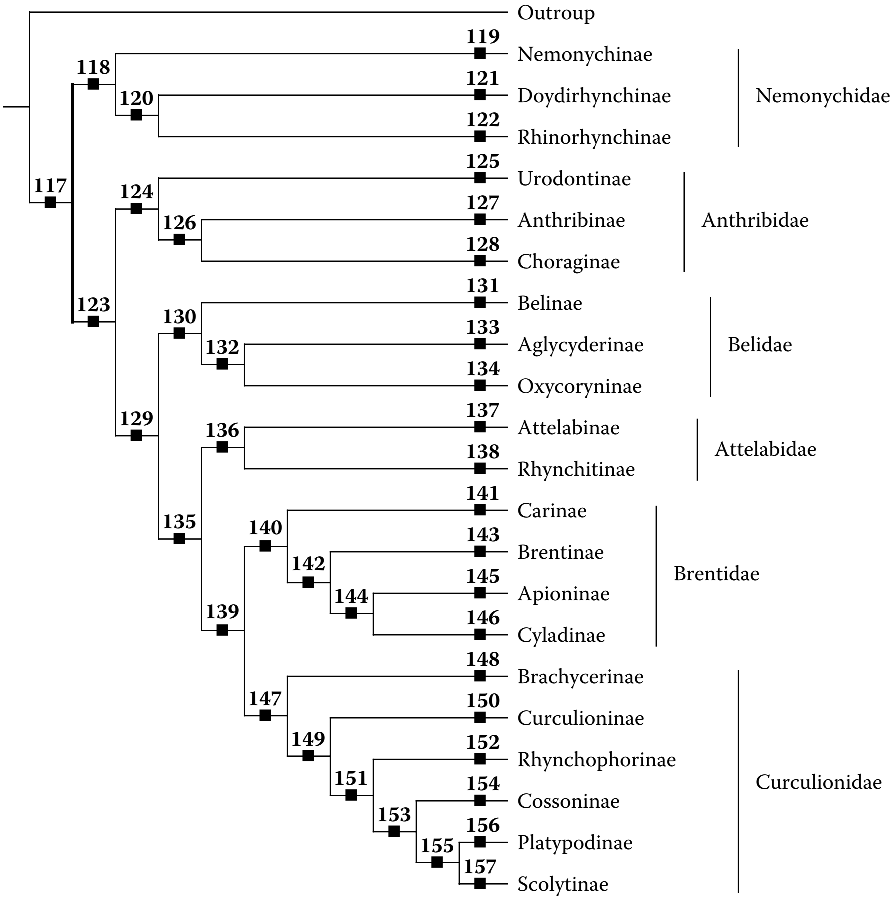


FIGURE 5.6 Phylogenetic classification of Curculionoidea sec. Kuschel (1995). Each concept is labelled with a unique number (see also Figure 5.8). Non-ranked concepts were assigned informal names–for example, concept 155 was named Platypodinae-Scolytinae sec. Kuschel (1995). The author introduced one new name (Carinae) in this system.

NAME/CONCEPT DISJUNCTION IN FIVE HIGHER LEVEL CLASSIFICATIONS OF WEEVILS

The preceding studies demonstrate that both the status and the meaning of Linnaean names continue to evolve from one authoritative revision to the next. What they cannot show very clearly, however, is the extent to which the transformations in naming and meaning become *disjunct* over time. For this purpose and also to complement the picture with a zoological example, five higher level classifications of weevils (Coleoptera: Curculionoidea) were analysed.

The classifications were authored by Crowson (1981), Thompson (1992), Kuschel (1995), Alonso-Zarazaga and Lyal (1999), and Marvaldi and Morrone (2000). Kuschel (1995) published the first matrix-based phylogeny for weevils, which was

TABLE 5.3
Quantitative Analysis of Relationships Linking Accepted Concepts in Weakley (2006) to Predecessors in Eight Pertinent Floras

Relationship Comparison	Relationship (%)					Nom./Tax. Stable[a]	
Weakley (2006) with ...	**==**	**>**	**<**	**><**	**|**	**%**	**Totals**
Kartesz (1999)	92.9	2.5	4.6	0.0	0.0	86.4	4064/4705
Flora of North America (1993)	93.9	0.5	5.6	0.0	0.0	87.5	1737/1985
Gleason and Cronquist (1991)	87.3	2.5	10.1	0.1	0.0	75.9	2385/3144
Godfrey and Wooten (1979, 1981)	82.4	1.1	16.4	0.0	0.0	72.8	975/1339
Radford et al. (1968)	81.1	2.6	16.3	0.0	0.0	68.7	1884/2742
Gleason (1952)	81.9	8.0	10.0	0.1	0.0	67.8	1866/2751
Fernald (1950)	77.1	16.4	6.2	0.3	0.0	63.5	1951/3073
Small (1933)	78.2	10.5	11.0	0.3	0.0	54.9	1571/2859

Note: Complete references in Weakley (2006).

[a] Nomenclature *and* taxonomy stable.

subsequently expanded and reanalysed by Marvaldi and Morrone (2000). The other three classifications are traditional (i.e. not cladistic). Alonso-Zarazaga and Lyal's (1999) *Catalogue* is the most recent comprehensive perspective on weevil taxonomy. The extent of topological variation among these perspectives is readily apparent (Figures 5.6 and 5.7).

A total of 172 names and 267 concepts were derived from the five taxonomies, and 1088 concept relationships were established among their constituent elements. The entire vocabulary for expressing relationships (Table 5.1) was used in order to maximize the amount of congruence between classifications. Comparisons that were labelled with a '>' or '<' simply because one system did not reach down to the same hierarchical level (i.e. inclusions per rank) were excluded from the analysis. The results are therefore as favorable towards the Linnaean system as possible with this data-set and approach.

The two-point comparisons of the five perspectives yielded only 18–54 per cent congruence among related concepts (Table 5.4). The numbers were expectedly lower when stability in naming *and* meaning was assessed, ranging from 6 to 29 per cent. In other words, each new treatment has made at least half of the preceding names and concepts unstable.

In all, 171 relationships were established between concepts carrying the same Linnaean name, and 597 relationships were made between concepts with different names (Table 5.5). These two sets of relationships are best suited to uncover the name/meaning disjunction inherent in the five taxonomies. Specifically, only 89 of the 171 nomenclaturally identical relationships (52.0%) were also taxonomically congruent. The other 82 relationships (48.0%) were either more or less inclusive, or overlapping. In each of these 82 cases the Linnaean names were unable to signal the changes in meaning. Overlap is typically the most complex kind of relationship; it

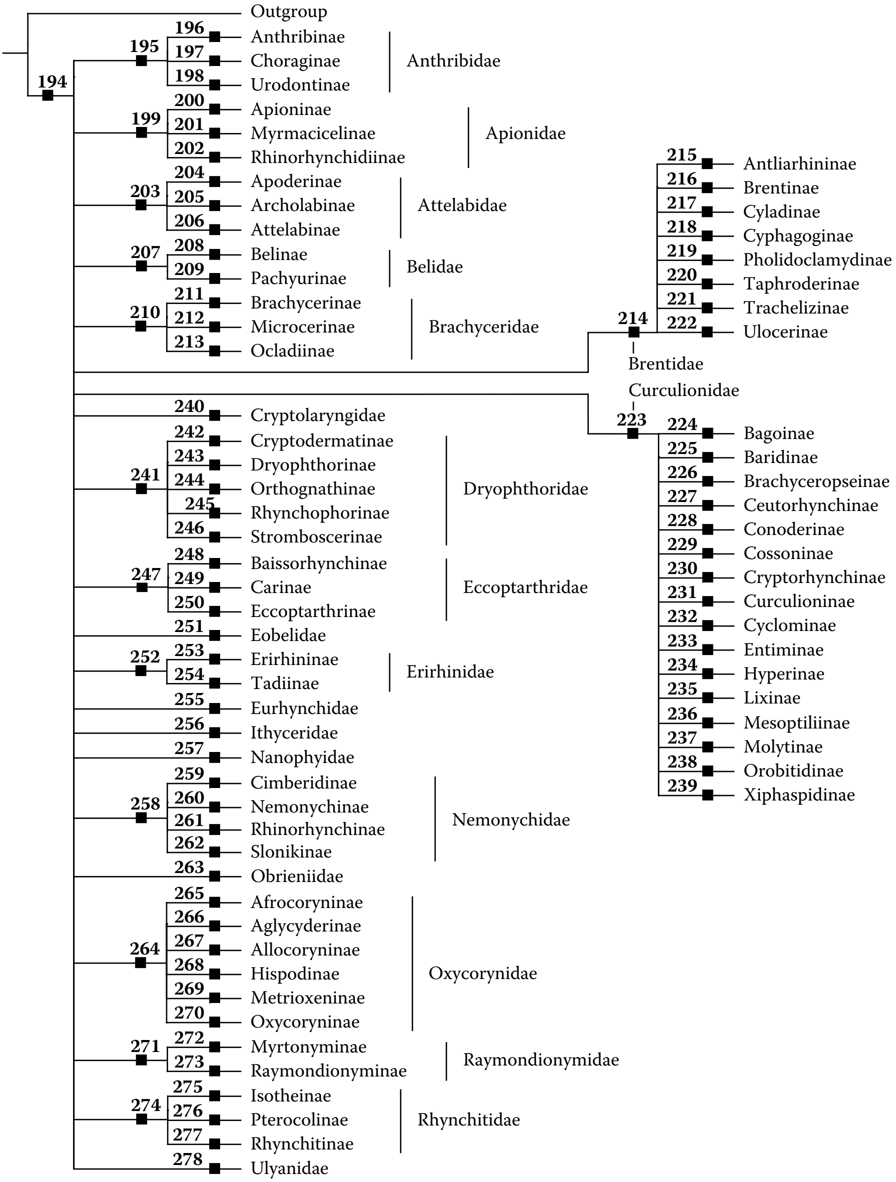


FIGURE 5.7 Classification of Curculionoidea (excepting Platypodidae and Scolytidae) sec. Alonso-Zarazaga and Lyal (1999), to the level of subfamily. Seven new names were proposed. To illustrate a concept relationship to Kuschel's (1995) system (Figure 5.6): Brentidae + Eurhynchidae – Cyladinae sec. Alonso-Zaraza & Lyal (1999) == Brentinae sec. Kuschel (1995). Note that only concept relationships are able to convey the inverse nestedness of elements in this example (i.e. a subfamily including a family).

TABLE 5.4
Quantitative Analysis of Relationships Linking Accepted Concepts in Five Succeeding Weevil Classifications to Each Other: Per Cent Values

Relationship Comparison[a]	Relationship (%)					Nom./Tax. Stable[b]	
Succeeding With Preceding Classification	**==**	**>**	**<**	**><**	**|**	**%**	**Totals**
M. and M. (2000) with A.-Z. and L. (1999)	38.7	24.0	18.7	4.0	14.7	6.7	5/75
M. and M. (2000) with Kuschel (1995)	26.8	25.0	26.8	21.4	0.0	12.5	7/56
M. and M. (2000) with Thompson (1992)	41.3	33.3	20.6	4.8	0.0	7.9	5/63
M. and M. (2000) with Crowson (1981)	18.2	34.5	34.5	12.7	0.0	10.9	6/55
A.-Z. and L. (1999) with Kuschel (1995)	33.7	9.9	44.6	4.0	7.9	12.9	13/101
A.-Z. and L. (1999) with Thompson (1992)	41.3	31.2	18.1	2.2	7.2	18.1	25/138
A.-Z. and L. (1999) with Crowson (1981)	30.8	9.6	40.4	5.8	13.5	5.8	3/52
Kuschel (1995) with Thompson (1992)	29.7	56.5	8.0	5.8	0.0	7.2	10/138
Kuschel (1995) with Crowson (1981)	37.1	25.8	22.6	14.5	0.0	11.3	7/62
Thompson (1992) with Crowson (1981)	53.6	14.3	32.1	0.0	0.0	28.6	8/28

[a] M. and M. (2000) = Marvaldi and Morrone (2000); A.-Z. and L. (1999) = Alonso-Zarazaga and Lyal (1999).

[b] Nomenclature *and* taxonomy stable.

means that the two classifications cannot be reconciled unless certain groups of subelements are added or subtracted from at least one side of the equation. The fact that there are no '|' relationships in the identical-name set is due to the method of types.

Within the other set where the compared names are not the same, 177 relationships (29.6%) are nevertheless taxonomically congruent (Table 5.5). Synonymy accounts for 13 of these comparisons (2.2%), whereas changes in rank–and thus in spelling–make up 30 additional cases (5.0%). The remaining 134 congruent relationships (22.4%) often represent very different nomenclatural perspectives, as illustrated in two examples of concept lineages for Brentidae and Curculionidae (Figure 5.8). Linnaean names are not capable of signaling the congruence in meaning in these cases. Among the 420 non-congruent relationships, the 47 assessments of overlap are also a sign of taxonomic complications (see previous discussion).

Not included in the analysis are variations in naming with purely nomenclatural origins. For instance, according to information from the *Catalogue* (Alonso-Zarazaga and Lyal, 1999; see also Figure 5.7), the 85 valid names are associated with 283 homotypic synonyms, 107 heterotypic synonyms and 155 names with incorrect spelling ('lapsus').[6] At least the homotypic synonyms and the misspelled names could in principle have come into existence without reexamining specimens or new taxonomic judgements. They might further promote the name/meaning disjunction.

In summary, quantitative analyses of concept evolution in German mosses, North American vascular plants, and weevils do not support the impression that Linnaean names are sufficiently precise to accommodate what researchers have learned throughout the decades about the relationships among these taxa. Instead,

TABLE 5.5
Quantitative Analysis of Relationships Linking Accepted Concepts in Five Succeeding Weevil Classifications to Each Other: Absolute Values and Name/Meaning Disjunction[b]

Relationship Comparison[a]	**Relationship**					**Total**
Nomenclature Stable in Comparison	==	>	<	><	|	
M. and M. (2000) with A.-Z. and L. (1999)	5	7	2	0	0	14
M. and M. (2000) with Kuschel (1995)	7	0	4	0	0	11
M. and M. (2000) with Thompson (1992)	5	4	0	1	0	10
M. and M. (2000) with Crowson (1981)	6	0	1	1	0	8
A.-Z. and L. (1999) with Kuschel (1995)	13	4	10	0	0	27
A.-Z. and L. (1999) with Thompson (1992)	25	11	8	0	0	44
A.-Z. and L. (1999) with Crowson (1981)	3	3	6	0	0	12
Kuschel (1995) with Thompson (1992)	10	11	1	2	0	24
Kuschel (1995) with Crowson (1981)	7	0	1	1	0	9
Thompson (1992) with Crowson (1981)	8	1	3	0	0	12
Total	**89**	**41!**	**36!**	**5!**	**0!**	**171**
Nomenclature Unstable in Comparison						
M. and M. (2000) with A.-Z. and L. (1999)	24	11	12	3	11	61
M. and M. (2000) with Kuschel (1995)	8	14	11	12	0	45
M. and M. (2000) with Thompson (1992)	21	17	13	2	0	53
M. and M. (2000) with Crowson (1981)	4	19	18	6	0	47
A.-Z. and L. (1999) with Kuschel (1995)	21	6	35	4	8	74
A.-Z. and L. (1999) with Thompson (1992)	32	32	17	3	10	94
A.-Z. and L. (1999) with Crowson (1981)	13	2	15	3	7	40
Kuschel (1995) with Thompson (1992)	31	67	10	6	0	114
Kuschel (1995) with Crowson (1981)	16	16	13	8	0	53
Thompson (1992) with Crowson (1981)	7	3	6	0	0	16
Total	**177!**	**187**	**150**	**47**	**36**	**597**

[a] M. and M. (2000) = Marvaldi and Morrone (2000); A.-Z. and L. (1999) = Alonso-Zarazaga and Lyal (1999).

[b] Marked in the totals with '!'; see text for further details.

the numbers demonstrate that the scenario described for the hypothetical taxon *Fantasia* (Figure 5.1) has abundant real-life parallels. Nomenclatural emendations and changes in taxonomic circumscription often evolve independently. Concept relationships provide the necessary resolution.

AUTHORITATIVE TAXONOMIC DATABASES—A PRIME APPLICATION FOR THE CONCEPT APPROACH

A more widespread adoption of the concept approach requires an efficient strategy for implementation. One area of application is the development and upkeep of

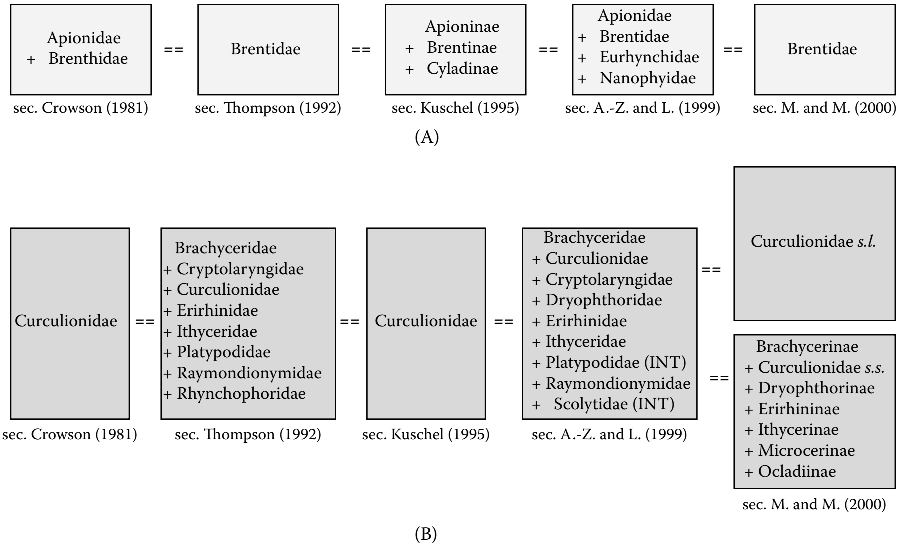


FIGURE 5.8 Two taxonomically congruent concept lineages including the names (A) Brentidae and (B) Curculionidae, as defined in five weevil classifications (Platypodidae and Scolytidae are not explicitly treated in Alonso-Zarazaga and Lyal [1999], thus in [INT] annotation). Two examples with single names are: Brentidae sec. Alonso-Zarazaga and Lyal (1999) >< Brenthidae sec. Crowson (1981); and Curculionidae *s.s.* sec. Marvaldi and Morrone (2000) > Curculionidae sec. Thompson (1992).

authoritative taxonomic databases (see also Berendsohn et al., 2003; Garrity and Lyons, 2003; Kennedy et al., 2005).[7] These databases are rapidly diversifying and have become indispensable tools for research. Examples are the USDA PLANTS Database (http://plants.usda.gov/), the BioSystematic Database of World Diptera (http://www.sel.barc.usda.gov/Diptera/biosys.htm), the Catalog of Fishes Online (http://www.calacademy.org/research/ichthyology/catalog/), and the Mammal Species of the World (www.nmnh.si.edu/msw/). The latter is based on a book with the same title published more than a decade ago (Wilson and Reeder, 1993). The names and classification proposed therein are routinely cited in mammal research.

Wilson and Reeder (personal communication) now have a completely revised version of the 1993 treatment. The new perspective contains significant changes in nomenclature and taxonomy; many are of the sort that cannot be expressed with names or synonymy relationships alone. In a name-based database, this all-too-common situation creates two almost equally undesirable options. The first option is to fully replace the old system with the new one. This would mean that the concepts advocated in 1993 are no longer available online. Consequently, other works in which these concepts were cited will lose their semantic underpinning. Users who assume that the older and newer names are taxonomically congruent incur the aforementioned risks of imprecision. The second option is to leave the database in its original state. But this amounts to a failure to adjust to latest and most supported perspectives. In other words, a name-based database system is unable to fully document its own taxonomic development.

The concept approach is well suited to overcome these challenges. Using the 'sec.' annotation, the 1993 and 2005 perspectives can both be displayed. Precise concept relationships would connect the elements contained in each taxonomy. Users can access this information to understand the proposed changes in meaning. The concept approach is also useful for occasional 'local' updates of particular taxa that have undergone revision after the latest comprehensive update went into print. Any attempt to capture the evolution of taxonomic perspectives in an online environment will in some form depend on this approach.

SCHEMAS AND TOOLS IN SUPPORT OF CONCEPT TAXONOMY

A 'taxonomic concept schema' has been created to promote the transition towards concept taxonomy (Hyam, 2005; Kennedy et al., 2005). The schema was written in XML and is based on an inclusive model for the representation and transfer of nomenclatural and taxonomic data. It accommodates a range of information stored in different formats without data distortion. The schema has been developed in close collaboration with the Taxonomic Databases Working Group community and was ratified as a standard for data transfer at the 2005 Annual Meeting in Saint Petersburg, Russia.[8] For providers interested in transforming their current holdings into concepts, the taxonomic concept schema will become an essential tool. In addition, there are numerous tools available that allow taxonomic experts to visualize two or more classifications and to infer or establish new concept relationships between their constituent elements (Graham et al., 2002; Güntsch et al., 2003; Munzner et al., 2003; Parr et al., 2004; Wang and Goguen, 2004; Liu et al., 2006). Such 'concept relationship tools' will combine the most powerful solutions for visualizing hierarchies with a full-scale implementation of concept taxonomy.

CONCLUSIONS–PROMISE AND PRACTICAL CHALLENGES FOR THE CONCEPT APPROACH

This chapter started out by describing the taxonomic resolution needs in a specific biodiversity research workflow. Linnaean names were shown to be too imprecise to support these needs, and taxonomic concepts and relationships were introduced as a more reliable long-term solution. This approach has so far been implemented with success in select taxonomic databases and regional floristic treatments. Quantitative analyses have added further weight to the claim that taxonomic concepts are suitable to overcome the problem of name/meaning disjunction. A full online documentation of the taxonomic process will therefore depend on a wider adoption of concept taxonomy. New tools are emerging towards this goal.

The concept approach improves communication about nature without compromising any of the useful properties of the Linnaean system. It does not aim to alter the method of types, the principle of priority, ranks or other nomenclatural rules and conventions–all of which play a critical role in making Linnaean names more precise and reliable.

It is worth reiterating that the added semantic granularity of concepts is not required in all contexts. In many everyday cases Linnaean names are precise enough or a considerable amount of vagueness is acceptable. In other situations human cognitive abilities come to assistance. Researchers who have been exposed to similar academic environments have amazing and often intractable capabilities to understand each other's uses of language. For instance, no living weevil taxonomist would think of the meaning of 'Derelomini' in the original mid-nineteenth century sense of the term. Instead, he or she will have in mind a list of the approximately 40 genera cited in the *Catalogue* (Alonso-Zarazaga and Lyal, 1999), complemented by mental images of examined specimens, and perhaps also an influential tribal definition for 'Petalochilinae' published by Kuschel (1952). A small group of experts will 'understand' that there are several unpublished problems with the position taken in the *Catalogue*. They may even have exchanged views about necessary changes, and so on. In other words, competent speakers are highly accustomed to using Linnaean names in reference to a specific published or unpublished context. Naturally, this is an implicit use of the concept approach. The challenge is to uncover this kind of implied precision and make it available to a wider audience.

The concept approach is furthermore an adequate response to the discussion about 'unitary taxonomy' (Scoble, 2004). Vane-Wright (2003) showed that it is almost impossible to arrive at a universally accepted classification for a particular taxonomic group. Working taxonomists tend to disagree not only with others but also with their *own* previous views. It could not be any other way if new evidence is supposed to count towards the meanings of scientific terms. Instead of forcing research (however authoritative) to a standstill, a more desirable benchmark for taxonomists is to precisely understand and document the *nature* of their disagreements. What they and other biodiversity researchers need first and foremost is the ability to *reconcile* the different views; this is what concept relationships will provide. Whether everybody uses exactly the same 'correct' taxonomy is neither as critical nor as realistic. In a close match with actual practice, the concept approach allows multiple competing taxonomic perspectives to coexist and gradually undergo refinement. It was invented by people with real-life data management and integration needs.

Lastly, a more widespread adoption of the concept approach will pose several challenges. The greatest among them is to minimize unnecessary 'concept inflation', or the proliferation of vaguely specified and potentially redundant concepts (Berendsohn, 1995). Indeed, in a world where the semantics of names are not fully defined unless their source is mentioned as well, every usage of a name must signal what its taxonomic source is. From a standpoint of effective communication, the ideal situation includes a pool of high-quality concepts that is only as large as necessary to accommodate all taxonomically diverging perspectives. The elements in the pool are connected to their closest matches via concept relationships. Users routinely cite these concepts in their publications. Taxonomic experts take a conservative approach towards authoring new concepts, preferring instead to credit a preexisting source whose perspective they accept (if such a match is available). In short, a successful implementation of the concept approach will require experts, providers, and users of taxonomic information to be very explicit about their speaker roles. What is the switch point going from authorship to citation of a concept? It will take time and intellectual as well as sociopolitical effort to adjust to this requirement in practice.

Another challenge is the integration of phylogenetic insights and traditional classifications. This challenge is not unique to the concept approach; however, the latter carries the highest promise of resolution (Franz, 2005). In modern systematics an increasing number of phylogenetic analyses are no longer translated into classifications, even though the precise transmission of phylogenetic insights depends on the frequent revision of Linnaean names. For those phylogeneticists who are typically not interested in classifying, the threshold will be lowered to author new concepts, without also having to author new names. They can therefore reach a wider audience with their products. But the realization of this prospect depends on a better physical and semantic integration of phylogenetic and taxonomic databases (see also Page, 2004, 2005, 2006).

To conclude, the taxonomic concept approach promises immense benefits for data integration in taxonomy, phylogenetics, and biodiversity research. The challenges related to implementation are considerable, yet in light of a community-wide motivation to ready taxonomy for the Semantic Web (Berners-Lee et al., 2001), it appears that time is on its side.

ACKNOWLEDGEMENTS

The authors would like to thank Quentin Wheeler for the opportunity to contribute this paper to the *New Taxonomy* session at the Fifth Biennial Meeting of the Systematics Association (National Museums and Galleries of Wales, Cardiff, 2005). Much of the presented material has been developed through discussion with members of the Science Environment for Ecological Knowledge (SEEK) Taxon Working Group. Mark Schildhauer provided helpful comments on earlier versions of the presentation and manuscript. This material is based upon work supported by the Andrew W. Mellon Foundation (postdoctoral training fellowship to NMF) and the National Science Foundation, under Award 0225676.

NOTES

1. See http://elib.cs.berkeley.edu/manis/.
2. The example could be modified to apply to higher level taxa such as families and genera or to characters instead of specimens.
3. The Australian Plant Name Index (http://www.anbg.gov.au/apni/) and the Euro+Med PlantBase (http://www.euromed.org.uk/) are two examples.
4. Note that the term ‘concept’ is not used here in the same sense as ‘species concepts’. Species concepts are theories about what species are, how they arise and how to recognize them (see Wheeler and Meier, 2000).
5. The possibility remains to connect taxonomic concepts via traditional nomenclatural relationships (homonymy, synonymy, etc.).
6. For quantitative analyses of rates of synonymization in a range of taxa see Olson (1987), Gaston and Mound (1993), Solow et al. (1995), Bouchet (1997) and Alroy (2002).

7. In the present context 'authoritative' means that the provided information was created according to standards that are very close to those established for a traditional publication in taxonomy.
8. See http://www.tdwg.org/.

REFERENCES

Alonso-Zarazaga, M.A. and Lyal, C.H.C. (1999) *A World Catalogue of Families and Genera of Curculionoidea (Insecta: Coleoptera) (Excepting Scolytidae and Platypodidae),* Entomopraxis, Barcelona, 315 pp.

Alroy, J. (2002) How many named species are valid? *Proceedings of the National Academy of Sciences,* 99: 3706–3711.

Berendsohn, W.G. (1995) The concept of 'potential taxa' in databases. *Taxon,* 44: 207–212.

Berendsohn, W.G. et al. (2003) The Berlin Model: A concept-based taxonomic information model. *Schriftenreihe für Vegetationskunde,* 39, 15–42.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The semantic web. *Scientific American,* May, 284: 34–43.

Bouchet, P. (1997) Inventorying the molluscan diversity of the world: What is our rate of progress? *The Veliger,* 40: 1–11.

Chapman, A.D. (2005) Principles and methods of data cleaning–Primary species and species occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen, 75 pp.

Crowson, R.A. (1981) *The Biology of the Coleoptera,* Academic Press, London, 802 pp.

Farber, P.L. (1976) The type concept in zoology in the first half of the nineteenth century. *Journal of the History of Biology,* 9: 93–119.

Franz, N.M. (2005) On the lack of good scientific reasons for the growing phylogeny/classification gap. *Cladistics,* 21: 495–500.

Franz, N.M. (2006) Towards a phylogenetic system of derelomine flower weevils (Coleoptera: Curculionidae). *Systematic Entomology,* 31: 220–267.

Garrity, G.M. and Lyons, C. (2003) Future-proofing biological nomenclature. *OMICS,* 7: 31–33.

Gaston, K.J. and Mound, L.A. (1993) Taxonomy, hypothesis testing and the biodiversity crisis. *Proceedings of the Royal Society of London, Series B,* 251: 139–142.

Geoffroy, M. and Berendsohn, W.G. (2003) The concept problem in taxonomy: Importance, components, approaches. *Schrifteneihe für Vegetationskunde,* 39: 5–14.

Graham, M., Watson, M.F. and Kennedy, J.D. (2002) Novel visualisation techniques for working with multiple, overlapping classification hierarchies. *Taxon,* 51: 351–358.

Güntsch, A. et al. (2003) The taxonomic editor. *Schrifteneihe für Vegetationskunde,* 39: 43–56.

Howell, A.H. (1901) Revision of the skunks of the genus *Chincha. North American Fauna,* 20: 8–63.

Hyam, R.D., ed (2005) *Taxon Concept Schema–User Guide, Version 1.0.* Available at http://tdwg.napier.ac.uk/TCS_1.0/docs/UserGuidev_1.0. pdf 28 pp.

Kennedy, J., Kukla, R. and Paterson, T. (2005) Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. In *Data Integration in the Life Sciences: Proceedings of the Second International Workshop* (eds B. Ludäscher and L. Raschid), San Diego, CA, July 20–22, DILS 2005, LNBI 3615, 2005, 80–95.

Koperski, M. et al. (2000) Referenzliste der Moose Deutschlands. *Schrifteneihe für Vegetationskunde,* 34: 1–519.

Kuschel, G. (1952) Los Curculionidae de la Cordillera Chileno-Argentina (1.^{a} parte) (Aporte 13 de Coleoptera Curculionidae). *Revista Chilena de Entomología,* 2: 229–279.

Kuschel, G. (1995) A phylogenetic classification of Curculionoidea to families and subfamilies. *Memoirs of the Entomological Society of Washington,* 14: 5–33.

Liu, X. et al. (2006) ConceptMapper: A new tool for establishing links between multiple taxonomic classifications. *ISEIS,* in press.

Ludäscher, B. et al. (2005) Managing scientific data: From data integration to scientific workflows. *GSA Today, Special Issue on Geoinformatics*, 21 pp.

Ludäscher, B. et al. (2006) Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows,* 18: 1039–1065.

Marvaldi, A.E. and Morrone, J.J. (2000) Phylogenetic systematics of weevils (Coleoptera: Curculionoidea): A reappraisal based on larval and adult morphology. *Insect Systematics & Evolution,* 31: 43–58.

Meier, R. and Dikow, T. (2004) The significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable and complete specimen data. *Conservation Biology,* 18: 478–488.

Michener, W.K. and Brunt, J.W. (2000) *Ecological Data: Design, Management, and Processing,* Blackwell Science, Malden, England, 180 pp.

Michener, W.K. et al. (1997) Non-geospatial metadata for ecology. *Ecological Applications,* 7: 330–342.

Michener, W.K. et al. (2007) A knowledge environment for the biodiversity and ecological sciences. *Journal of Intelligent Information Systems,* 29: 11–126.

Munzner, T. et al. (2003) TreeJuxtaposer: Scalable tree comparison using Focus+Context with guaranteed visibility. *ACM Transactions on Graphics,* 22: 453–462.

Olson, S.L. (1987) On the extent and source of instability in avian nomenclature, as exemplified by North American birds. *Auk,* 104: 538–542.

Page, R.D.M. (2004) Towards a taxonomically intelligent phylogenetic database. *Technical Reports in Taxonomy,* 04-01, 1–5. Available at http://taxonomy.zoology.gla.ac.uk/publications/tech-reports/Edinburgh.pdf 5 pp.

Page, R.D.M. (2005) Phyloinformatics: Towards a phylogenetic database. In *Data Mining in Bioinformatics, Advanced Information and Knowledge Processing,* Vol. XI (eds M.J. Zaki, H.T.T. Toivonen, J.T.L. Wang and D.E. Shasha), Springer Verlag, Berlin, 219–241.

Page, R.D.M. (2006) Taxonomic names, metadata, and the Semantic Web. *Biodiversity Informatics,* 3: 1–15.

Parr, C.S., Lee, B., Campbell, D. and Bederson, B.B. (2004) Tree visualizations for taxonomies and phylogenies. *Bioinformatics,* 20: 2997–3004.

Paskin, N. (2005) Digital object identifiers for scientific data. *Data Science Journal,* 4: 12–20.

Peterson, A.T. and Navarro-Sigüenza, A.G. (1999) Alternate species concepts as bases for determining priority conservation areas. *Conservation Biology,* 13: 427–431.

Peterson, A.T. et al. (2002) Future projections for Mexican faunas under global climate change scenarios. *Nature,* 416: 626–629.

Scoble, M.J. (2004) Unitary or unified taxonomy? *Philosophical Transactions of the Royal Society of London, Series B,* 359: 699–711.

Soberón, J. and Peterson, A.T. (2004) Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London, Series B,* 359: 689–698.

Solow, A.R., Mound L.A. and Gaston, K.J. (1995) Estimating the rate of synonymy. *Systematic Biology,* 44: 93–96.

Stevens, P.F. (1984) Metaphors and typology in the development of botanical systematics 1690–1960, or the art of putting new wine in old bottles. *Taxon,* 33: 169–211.

Thompson, R.T. (1992) Observations on the morphology and classification of weevils (Coleoptera, Curculionoidea) with a key to major groups. *Journal of Natural History,* 26: 835–891.

Vane-Wright, R.I. (2003) Indifferent philosophy versus almighty authority: On consistency, consensus and unitary taxonomy. *Systematics and Biodiversity,* 1: 3–11.

Wang, G. and Goguen, J. (2004) *Analysis for Schema Matching Tool User Interface Design,* Technical Report, Department of Computer and Engineering, UCSD. Available at http://www.cse.ucsd.edu/users/guilian/reports/ui_analysis.pdf 14 pp.

Weakley, A.S. (2006) Flora of the Carolinas, Virginia, and Georgia. Working draft of January 17, 2006. Available at http://www.herbarium.unc.edu/flora.htm 1015 pp.

Wheeler, Q.D. (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society of London, Series B,* 359: 571–583.

Wheeler, Q.D. and Meier, R., eds (2000) *Species Concepts and Phylogenetic Theory: A Debate,* Columbia University Press, New York, 230 pp.

Wilson, D.E. and Reeder, D.M., eds (1993) *Mammal Species of the World,* Smithsonian Institution Press, Washington, D.C., 1207 pp.

6 International Infrastructure for Enabling the New Taxonomy

The Role of the Global Biodiversity Information Facility (GBIF)

Larry Speers and James L. Edwards

CONTENTS

INTRODUCTION

Taxonomy is dependent upon access to literature, specimens and data. Because many such necessary resources for taxonomy are located in the developed world and much of the undescribed biodiversity exists in the developing world, there exists a need to mobilize collections-based information and make it openly available. Existing and developing programmes of the Global Biodiversity Information Facility (GBIF) are described that are overcoming these barriers and speeding the advance of taxonomy.

Numerous authors have documented the urgency of increasing our understanding of our planet's biodiversity. It has been estimated that there are probably between 4 and 10 million species that cohabit our planet, of which approximately 1.75 million have been described scientifically. The major challenge for the New Taxonomy will be the discovery and description of the remaining undescribed taxa. For this task to be completed in a timely manner the current rate of species description will have to be increased many fold. If we hope to obtain this rate of increase, the existing bottlenecks and inefficiencies in current taxonomic practices need to be reviewed and where possible solutions identified that will help accelerate the process.

In general, working taxonomists describe new taxa by determining that certain specimens are outside the circumscription of previously described taxa. This process ideally entails:

1. Reviewing all relevant species descriptions in the available taxonomic literature.
2. Comparing these literature-based species descriptions with characters and character states of as large a series of closely related specimens as possible. Usually, the specimens that are examined have been obtained by:
   a. selecting appropriate specimens from the collection at the researcher's home institution;
   b. obtaining similar specimens through new collecting efforts;
   c. visiting other institutions to examine their holdings for similar material; or
   d. requesting the loan of relevant material from the collections visited or other institutions.
3. Validating, where possible, the literature descriptions by examining all relevant type material, either through loans or by visiting the institutions holding this material.

Once particular specimens or series of specimens have been identified as being outside the circumscription of all currently named taxa, the researcher can formally name this material as a new taxon by designating appropriate type material for future reference and submitting this new taxonomic description to the scientific community through the peer-reviewed publication process.

The basic process for describing new taxa has changed little since it was originally formalized through the adoption of the original Codes of Nomenclature over a century ago. Each of the steps in this process needs to be examined to see if there are inherent inefficiencies that can be addressed in light of recent advancements in information technologies and modern approaches to information management.

THE TAXONOMIC LITERATURE

Unlike most other sciences, the science of taxonomy is totally dependent on 200 years of legacy literature (Minelli, 2003), and comparing the original taxonomic descriptions found in this literature with specimens is a critical step in the taxonomic process. Unfortunately, the taxonomic literature is very scattered and there is no authoritative index to this vast resource. Species descriptions have been published in many languages and some of the literature, particularly the older literature, is of varying degrees of quality. In addition, a lot of critical information is published in the 'grey literature', with limited circulation. Unless one is fortunate enough to have access to one of the few libraries in the world with significant taxonomic holdings, accessing much of this historical information through interlibrary loans is a slow and relatively expensive process. Most taxonomists spend significant time and resources throughout their careers doing the literature-based research that is necessary to build their own reference collections related to their taxonomic specialties. The accessing and assembling of this historical literature resource is a particular problem for workers in developing countries, who often do not have easy access to the larger library networks that are found in the developed world. It is important to note that most of the world's taxonomic literature resources are found in the developed world, while most of the megadiverse biota that contains the majority of undescribed taxa is found in the developing world. This mismatch of resources is currently a major obstacle to taxonomic research in developing countries.

Utilizing modern information technologies to create, index and distribute electronic versions of the taxonomic literature, particularly the older literature, would make possible more rapid and comprehensive access and greatly increase the efficiency of this step in the taxonomic process. Since the taxonomic process is so dependent on access to literature resources, any increase in efficiency in this area would accelerate the rate and improve the quality of taxonomic research. Having rapid electronic access to the taxonomic literature would also greatly enhance the capacity of taxonomic researchers in the developing world.

FINDING AND GAINING ACCESS TO SPECIMENS

The second step in the descriptive process involves comparing the characteristics of voucher specimens with the species descriptions found in the literature. Unfortunately, few collections have inventories that provide detailed information on the taxonomic, geographical and temporal distribution of their holdings. Without such a detailed inventory it is very difficult to predict what might be found in any collection and in most cases the only practical way for a researcher to locate relevant voucher material outside his or her home institution is to physically visit as many collections as possible and search through their holdings. The taxonomic, geographical and temporal distribution of the holdings of each collection is a direct result of the opportunistic way these holdings have expanded through time.

At any particular period of time in its development, the growth of any natural history collection has been dependent on the particular taxonomic and geographical interests of its curatorial and research staff and their access to resources to support collecting and curation activities. As a result, the taxonomic emphasis and geographical source of new material being added to any collection has changed through time as staff and their interests changed, as funding agencies changed their emphases and as the geopolitical landscape changed. For example, older material from East Africa is often located in German collections, while more recent material is to be found in British collections and even newer material to be found in collections in countries such as Denmark and Finland that support current research programmes in East Africa through their foreign aid programmes.

Due to the almost haphazard way the holdings of collections have accumulated, not only is it difficult to predict the holdings of any collection, but also each collection usually holds only a small portion of the material that should be examined to complete a comprehensive taxonomic revision. This undocumented and scattered distribution of critical material forces taxonomists to spend a significant proportion of their resources travelling the world, visiting various collections in search of additional material that will help validate their conclusions. Unfortunately, since it is possible only to visit a very small proportion of all collections, a vast amount of potentially relevant material is never examined and much of the long-term investment in collecting, storing and curating this unexamined material is wasted. One of the most significant benefits of databasing a collection is the discovery and documentation of its holdings from taxonomic, geospatial and temporal perspectives (Peterson and Navarro-Sigüenza, 2003).

The current emphasis in this step of the taxonomic process on moving people (taxonomists) and material (specimens through loans) rather than the more cost-effective and efficient process of moving information could be radically modified if the holdings of the world's natural history collections were databased and indexed and this information shared through openly accessible interoperable networks. Sharing this information would result in a far more efficient utilization of existing resources and facilitate taxonomists accessing a much higher percentage of the information related to specimens important for their research. In addition, since most type specimens and the majority of the world's natural history collections are located in the developed world, global access to specimen-based information would allow taxonomists in the developing world to far more efficiently contribute to the global taxonomic effort.

A common problem, in comparison with today's standards, is that much of the historical material in the natural history collections is inadequately labelled, particularly in relation to collection locality information. An additional benefit of databasing this material would be that it would allow the reconnection of material collected during a single collecting event that is now distributed across multiple collections. This would allow the georeferencing of these duplicate localities to be done only once with this information then shared between collections, rather than the current inefficient procedures where each specimen is georeferenced independently. Also, sharing information between collections will facilitate the identification of collection gaps–both taxonomically and geographically–and will make the information associated with our existing specimens much more accessible to additional user groups, such as ecological modellers, molecular phylogeneticists, etc.

Type specimens support the stabilization of nomenclature (Daston, 2004) and the examination of designated type material is critical to determining nomenclatural relationships and the circumscription of taxonomic concepts. Taxonomists currently spend significant time and resources determining where existing types are located and obtaining access to this material. Having an online index to type material including where it is located would greatly increase the efficiency of the taxonomic process.

Considering how critical type material is to the taxonomic process it is surprising that none of the nomenclatural codes specifically requires that this critical material be deposited in public institutions (Knapp et al., 2004). In some groups, between 30 and 60 per cent of type material could be deemed as lost or, because it is held in private collections, not available for study (Nash, 2001). Ensuring that any newly designated type material is properly documented and deposited in an institution with a long-term mandate to ensure its maintenance and accessibility for future researchers would minimize these problems.

Since type material is irreplaceable, collections holding these specimens are becoming increasingly reluctant to loan this material. As a result, rather than being able to borrow type specimens, taxonomists are increasingly being forced to travel to type-holding institutions to examine this material on site. Often researchers discover after investing in these visits that the material they have travelled to examine is not really critical to their current research. Recent advances in imaging technology and increased Internet connectivity have made it relatively inexpensive to share high-quality photographs or E-types (Speers, 2005). While access to these E-types will not eliminate the need to physically examine critical material, it could greatly reduce the number of types that need to be examined and as a result the number of institutions that need to be visited.

DOCUMENTING AND SHARING CHARACTER AND CHARACTER STATE INFORMATION

As a result of the constraints imposed by the limitations of the paper-based media on sharing large amounts of character and character state information, a significant amount of data collected on individual specimens is never made available for future consideration. This often means that subsequent researchers are forced to repeat the some measurements or score the same characters. Storing this information and making it digitally accessible would greatly accelerate future analysis.

THE ROLE OF GBIF IN SPEEDING UP THE TAXONOMIC PROCESS

The Global Biodiversity Information Facility (GBIF) (www.gbif.org) was established in 2001 with the goal of making primary biodiversity data freely and universally available over the Internet. GBIF acts as a broker to point taxonomists and other users to a wide array of digitized biodiversity resources. With its many partners and collaborators, GBIF can thus play a crucial role in increasing the efficiency of the taxonomic process, through activities and services such as the following.

Providing Tools for Searching Interoperable Databases and Sharing Taxonomic Data

Expanding on the distributed database architecture first developed by FishNet, (http://fishnet2.net/index.html) and the Mammal Networked Information System (MaNIS, http://manisnet.org/), as of December 2007, GBIF's data portal provided access to more than 141 million specimen and observation records served by 214 data providers. Currently, these records can be searched in a variety of ways. By early 2007, many additional capabilities will be implemented, including the following:

- Explore species: Find information organized by species, including names, classification, distribution, images and links to further resources.
- Explore countries: Find information organized by country, ecoregion or ocean area, including species known from the area, data providers located in the area and data resources about the area.
- Explore D: Find information organized by data provider or data resource, including information on data usage, species and countries included in each resource.
- Explore O: Find and map specimen and observational records.
- Advanced S: The user can combine any of the preceding items.

Digitizing and Georeferencing of Voucher Specimens

Since 2003 the GBIF programme on Digitization of Natural History Collections (DIGIT) has partnered with more than 40 institutions in more than 30 countries to database and share the label information associated with more than four million specimens and observational records. In total approximately USD 8 has been invested in this activity.

Imaging of Types and Other Key Specimens

In addition, the DIGIT seed-money programme has partnered with a number of institutions to image and share the images of more than 70,000 type specimens. GBIF has also partnered with the European Network for Biodiversity Information to produce a manual of best practices for digital imaging of biological type specimens (Häuser et al., 2005).

Developing a Computerized Reference Taxonomy

To facilitate access to taxonomic information the GBIF Electronic Catalogue of Names of Known Organisms (ECAT) work programme has been partnering with the Catalogue of Life partnership (http://www.catalogueoflife.org/search.php), uBio (http://www.ubio.org/) and others to produce, by 2011, an electronic catalogue of all of the scientific names of organisms that have ever been published, as well as a significant number of vernacular names.

Providing Documentation and Software Tools for Efficiently Improving Data Quality

Assessment of data quality is critical to determining whether data are 'fit for use' in any particular application. GBIF has made available two documents (Chapman, 2005a,b) to help the community better understand aspects of data quality and facilitate the improvement of data quality in the records shared through the GBIF network. In addition, with support from GBIF and the Gordon and Betty Moore Foundation, the Centro de Referência em Informação Ambiental (http://www.cria.org.br/) has developed a set of free, open-source software tools to simplify the task of checking a set of biodiversity data records for accuracy. These tools are available for download (https://sourceforge.net/project/showfiles.php?group_id=103853&package_id=165541).

Developing Globally Unique Identifiers for Collections, Specimens and Taxonomic Concepts

A globally unique identifier (GUID) is a permanent computerized tag used to identify, reference and retrieve data resources (i.e. information) through the Internet. A GUID framework for biodiversity that uniquely and permanently identifies individual specimens, collections, publications, etc. will be an invaluable tool for managing and cross-linking the many different types of entities that are basic to taxonomic research. In addition, GUIDs will greatly improve interoperability with other life sciences domains, such as molecular informatics and ecology. With support from the Gordon and Betty Moore Foundation, the Taxonomic Databases Working Group (TDWG) (http://www.tdwg.org/) and GBIF have initiated a broad community-based discussion to identify the requirements for globally unique identifiers to support biodiversity informatics (http://wiki.gbif.org/guidwiki/wikka.php?wakka=GUID1Report) and the New Taxonomy.

Providing Access to Taxonomic Literature

Although GBIF currently does not provide systematic access to taxonomic literature, developing methodologies for linking to that literature is one of the major goals of GBIF's current strategic plan. With the impending availability of large-scale literature digitization projects, such as the Biodiversity Heritage Library (http://www.biodiversitylibrary.org), it is imperative that GBIF be able to provide links between specimen information (especially the information for type specimens) and literature descriptions. The GUIDs described previously are a key component in developing unambiguous linkages among specimens, literature descriptions and taxon concepts.

Training the Next Generation of Taxonomists and Other Users of Digitized Information, Especially Individuals from the Developing World

The GBIF and its partners recognize that the global capacity to access and utilize information is very uneven. Through its work programme for Outreach and Capacity Building (OCB), GBIF aims to provide software tools and training to bridge biodiversity information technology gaps for all countries around the world. GBIF also addresses scientific and technical collaboration in many areas, including repatriation of data and intellectual property rights. As a result, GBIF is a major contributor to the Global Taxonomy Initiative (http://www.cbtint/gti/about.shtml) of the Convention on Biological Diversity.

CONCLUSIONS

In the future it is expected that the major impediment to taxonomic research will be the collecting and processing of new material to complete our knowledge base (May, 2004). It is important as we move into this phase that the problems we have identified in working with our existing legacy material are mitigated by ensuring that the information associated with this new material is properly documented and that this information flows seamlessly into the emerging taxonomic information infrastructure. The most significant return on investment that can currently be achieved is to also ensure that newly collected specimens, newly curated material, and newly published information are all made electronically accessible as part of the curation, research and publishing process. There is an urgent need for funding agencies and institutions that support taxonomic research to review their current policies concerning information management to ensure that future taxonomic research is not burdened with the same inefficiencies that impede present-day researchers.

REFERENCES

Chapman, A.D. (2005a) Principles of data quality. Report for the Global Biodiversity Information Facility, Copenhagen (available at http://www.gbif.org/prog/digit/data_quality).

Chapman, A.D. (2005b) Principles and methods of data cleaning–Primary species and species-occurrence data. Report for the Global Biodiversity Information Facility, Copenhagen (available at http://www.gbif.org/prog/digit/data_quality).

Daston, L. (2004) Type specimens and scientific memory. *Critical Inquiry,* 31: 153–182.

Häuser, C.L., Steiner, A., Holstein, J. and Scoble, M.J. (2005) *Digital Imaging of Biological Type Specimens. A Manual of Best Practice.* Results from a study of the European Network for Biodiversity Information. Staatliches Museum für Naturkunde, Stuttgart, viii + 304 pp.

Knapp, S., Lamas, G., Lughadha, E.N. and Novarino, G. (2004) Stability or stasis in the names of organisms: The evolving codes of nomenclature. *Philosophical Transactions of the Royal Society of London,* 359: 611–622.

May, R.M. (2004) Tomorrow's taxonomy: Collecting new species in the field will remain the rate-limiting step. *Philosophical Transactions of the Royal Society of London B,* 359: 733–734.

Minelli, A. (2003) The status of taxonomic literature. *Trends Ecology Evolution,* 18: 75–76.

Nash, R. (2001) Insect collections: A strategy for databases and acquisition. In *Biological Collections and Biodiversity* (eds B.S. Rushton, P. Hackney and C.R. Tyrie), Westbury Academic & Scientific Publishing, Otley, England, pp. 299–304.

Peterson, A.T. and Navarro-Sigüenza, A.G. (2003) Computerizing bird collections and sharing collection data openly: Why bother? *Bonner Zoologische Beiträge,* 51: 205–212.

Speers, L. (2005) E-types–A new resource for taxonomic research. In *Digital Imaging of Biological Type Specimens, a Manual of Best Practice. Results from a Study of the European Network for Biodiversity Information* (eds C.L. Häuser, A. Steiner, J. Holstein and M.J. Scoble), Staatliches Museum für Naturkunde, Stuttgart, pp. 13–18.

7 DNA Sequences in Taxonomy

Opportunities and Challenges

Rudolf Meier

CONTENTS

INTRODUCTION

In this chapter I discuss how DNA sequences can be used to solve existing taxonomic problems, discover new species, identify biological tissues, associate different life history stages and modernize the image of taxonomy in an attempt to attract 'big science' funding. However, I also join other authors in arguing that any taxonomy relying too heavily on DNA sequences is flawed. Given these problems, I argue for an integrative taxonomy which may help in unifying the taxonomic community in an attempt to attract 'big science' funding for the ambitious task of describing all of our planet's species.

Traditional taxonomy is in crisis. The literature is often inaccessible, electronic data management is in its infancy, retiring taxonomists are leaving numerous 'orphan' taxa behind, few students are entering the field, fewer yet are getting hired and academic administrations focusing on impact factors are abandoning the field altogether (Godfray, 2002; Pennisi, 2003a; Godfray and Knapp, 2004; Wheeler, 2004). Yet, taxonomists are more urgently needed than ever, given that 90 per cent of all species remain undescribed (Wilson, 2003), environmental impact assessments require accurate taxonomic information (Gaston and O'Neill, 2004) and conservation managers critically need precise species data (Sodhi et al., 2004). However, at the current rate traditional taxonomy will require more than 940 years before all species will be described (Seberg, 2004) because the doubling rate for species descriptions is quickly slowing (Wheeler, 2004). Under these circumstances, it does not come as a surprise that biologists are looking for new techniques that can speed up the process of describing species and identifying specimens. It is also not surprising that these biologists are looking at molecular techniques because they have successfully revitalized and significantly improved our understanding in many fields of biology.

Two radical proposals by Tautz et al. (2003) and Hebert et al. (2003a) have attracted considerable attention. Both assign larger roles to the use of DNA sequences in taxonomy. Given the importance of taxonomy it is not surprising that these proposals have received much attention. Hebert et al. (2003a) proposed 'DNA barcoding', which is now pursued by the Consortium for the Barcode of Life (CBOL). In a barcoded world, specimen identification will be based on a partial DNA sequence for COI. Investigators will identify an unidentified specimen by first extracting its DNA, then amplifying and sequencing COI before comparing the sequence from the query with COI sequences for all known species. Although more modest in language, Tautz et al. (2003) proposed an even more central role for DNA sequences in taxonomy when they envisioned a 'DNA taxonomy' using DNA sequences as a scaffold; i.e., not only specimen identification but also the determination of species boundaries and hence species descriptions would be based on DNA sequences. In this chapter, I will illustrate how DNA sequences have already contributed to taxonomy, discuss shortcomings of current approaches and finish with an outlook on the future use of sequences in taxonomy.

A BRIEF HISTORY

When reading the recent literature on DNA taxonomy and DNA barcoding, the reader may get the impression that it was Tautz et al. (2003) and Hebert et al. (2003a) who first proposed the use of DNA sequences for taxonomic purposes. However, this impression is far from being accurate (Sperling, 2003). For several decades, DNA sequences have played an important role in systematics. In the 1990s, many studies that were originally designed to reveal phylogenetic relationships between taxa, but the same studies often also yielded information on the genetic diversity below the species level. The study of intraspecific genetic variability quickly became a respected area of inquiry with several distinct historical roots.

First, systematists with classical training started to collaborate with molecular biologists in order to address species limits in particularly difficult taxa (e.g. Frost et al., 1998). DNA sequences were often able to provide the additional information that was needed to decide on species boundaries and/or the status of a population as a species or subspecies. Second, systematists with training in molecular phylogenetics started to use DNA sequences for studying species-level phenomena and even argued for sequencing standardized markers across a wide range of taxa in order to develop a database that could be used for cross-referencing (Caterino et al., 2000). Third, population biologists who had long studied gene flow within and between populations using a variety of techniques ranging from allozymes to DNA hybridization switched to DNA sequences once the sequencing became more widely available and affordable. This was the main origin of ‘phylogeography’, a field designed to address biogeographical relationships within species and species groups (Avise, 2000). Much of this research revealed higher than expected genetic diversity within some recognized species and the term ‘cryptic species’ emerged as a way to express the opinion that a traditional species contained multiple species (Bickford et al., 2007).

That DNA sequences can also be used to identify biological tissues to species (DNA barcoding) was similarly not a new proposal. For example, in 1998 Palumbi and Cipriano (1998) had used DNA sequences from whale meat bought in Asia to demonstrate that endangered species were sold and Birstein et al. (1998) investigated which sturgeon species were the source of the caviar sold in New York City. Even more widespread was the use of DNA sequences for identifying mosquitoes (Krzywinski and Besansky, 2003) and associating different life history stages (Paquin and Hedin, 2004; Steinke et al., 2005; Vences et al., 2005a). These are just a few examples and many additional cases can be found in the literature. I can thus conclude that overall the identification of biological tissues based on DNA sequences already had an illustrious history by the time Hebert et al.’s (2003a) proposal was published.

Given this extensive history of using DNA sequences in taxonomy, one can only be amazed about the response in the scientific and the popular press to Hebert et al.’s (2003a) proposal of DNA barcoding. Tautz et al.’s (2003) proposal of a DNA taxonomy similarly received more attention than could have been expected given that similar research had been carried out for at least a decade. However, what both proposals shared was a bold vision. DNA barcoding promised to make one of the hardest and least glamorous tasks in biology–the identification of specimens–easy, and Tautz et al. (2003) proposed a paradigm shift away from a largely morphology-based to a largely DNA sequence-based taxonomy. DNA barcoding had the additional appeal that it was promising to be 'big science' with the kind of budgets unheard of in taxonomy and that it was attractive to molecular biologists who are constantly sourcing for new funding. Both proposals also promised quick fixes to a long-standing problem that is starting to affect an increasingly large number of biologists: the crisis in taxonomy. This crisis had been recognized by scientists and politicians alike, but the solutions proposed by taxonomists had not been very palatable in being very expensive and slow.

DNA SEQUENCES IN TAXONOMY: OPPORTUNITIES

DNA Sequences Contribute New Data for Resolving Existing Taxonomic Problems

Resolving species boundaries between closely related species is notoriously difficult. Such species can generally only be diagnosed based on few characters that often have a host of problems. For example, the morphological differences between the species may be very subtle, difficult to describe and applicable only to some life history stages or only one gender. In other cases, taxonomists may encounter populations that are polymorphic with regard to characters that are normally diagnostic for species. This will leave him wondering whether he is dealing with one or multiple species. These problems are particularly pronounced in those cases where too few specimens are available for properly assessing the variability within and between populations. It is here that DNA sequences are an obvious and very rich source of additional data. The following recent cases from my laboratory may document these problems and potential solutions based on DNA sequences.

Ornithomya louseflies are obligate bloodsuckers on birds. The females are remarkable in larvipositing one mature larva at a time, which will pupate within a few hours of deposition. Four species occur in north-western Europe and all have been studied in detail in Scandinavia (Hill et al., 1964) and Britain (Hill, 1962b, 1963). However, the morphological work could not conclusively resolve whether *Ornithomya fringillina* (Curtis, 1836) and *O. chloropus* (Bergroth, 1901) were conspecific (Johnsen, 1948; Bequaert, 1953, 1954;) or not (Hill, 1962a; Hill et al., 1964; Theodor and Oldroyd, 1964; Hutson, 1984). The habitus of the two nominal species is almost identical, but four morphological and one life-history character had been proposed for separation. These were (a) wing length (3.5–4.5 mm vs. 4.5–5.5 mm), (b) number of scutellar bristles (four vs. six), (c) size of a dark spot on the ventral side of the head (absent or small vs. present and reaching jugular bristles), (d) degree of setosity on the wing (less setose vs. more setose: Hill et al., 1964; Hutson, 1984), and (e) duration of pupal stage (271 vs. 371 days; Hill, 1963).

Recently, we collected 13 specimens from the area of sympatry for the two nominal species and checked whether the morphological characters allowed for an unambiguous identification (Petersen et al., 2007a). One specimen was immediately identifiable as *O. fringillina* and two as *O. chloropus* based on all four morphological characters. The remaining 10 specimens could only be determined using a subset of the characters. Only the dark spot on the head was consistently useful. In conclusion, this case had many of the hallmarks of a species boundary problem caused by somewhat inclusive morphological characters: Some characters showed continuous variation, others were virtually inaccessible (pupation time) and the remaining one was difficult to define and had the additional problem of fading in old museum specimens (head marking).

Overall, the morphological evidence suggested the presence of two species based on one character. However, given the nature of the character, this conclusion was weak and crying out for more data. Given the extremely reduced nature of male genitalia in *Ornithomya,* DNA sequences were the obvious source. We sequenced COI for all 13 specimens collected in the area of sympatry and found that the genetic variability within *O. fringillina* and within *O. chloropus* was low (<0.5%) while the interspecific variability was high (>8%). The sequences thus provided strong additional evidence that the two forms are indeed two different species and we were since able to show that they are not even sister species within *Ornithomya* (Petersen et al., 2007b). Clearly, DNA sequences here proved useful as additional data for confirming species boundaries that remained poorly supported based on the traditional evidence.

A second case from my laboratory involved pentatomid bugs in the genus *Halys* (Memon et al., 2006). When Memon inspected the male and female genitalia of several series of specimens collected in Shahdadkot, Tandojam and Hyderabad (Sindh, Pakistan), she realized that characters that are usually constant within pentatomid species were rather variable within what appeared to be one species based on all other characters. The morphological variability did not fall into discrete clusters that would have suggested the presence of several species. However, relatively few specimens were available and thus more data were needed to confirm this result based on morphology. Memon et al. sequenced COI for those specimens that had the puzzling morphological differences and found that the genetic variability was negligible (<0.4%; Memon et al., 2006). Yet, the sequences from the same specimens were distinctly different from a closely related species that had been collected on the same host plant. Again, DNA data were able to shed light on a species boundary issue that was difficult to resolve based on traditional characters.

DNA Sequences Discover New Species

The literature is full of examples of widespread species composed of populations with an unusually large amount of genetic distinctness. Some of these populations are probably new species that have escaped discovery during the normal taxonomic work. This is particularly common in those groups that have a morphology that is difficult to study or lacks many features (e.g. protists, prokaryotes). But the same phenomenon is also not uncommon in morphologically complex multicellular animals. I believe that in most cases it is not the traditional taxonomic techniques *per se* that are to be blamed (Scotland et al., 2003). Many taxa have not been subjected to a taxonomic revision using modern techniques in many decades and it is only natural that old revisions will include lumped species because fewer and different microscopic techniques were used in the past. For example, in insects it was not until entomologists started to study genitalia that many new species were discovered and it was not until microscopial techniques became available that even nematodes were found to 'harbour endless morphological diversity' (De Ley et al., 2005, p. 1946). In other taxa, such as birds and frogs, detailed analyses of call data have similarly yielded evidence for multiple species where only a single species had been suspected based on morphology. In yet other cases the discovery of putatively new species based on DNA sequences can be attributed to new sampling. Due to the fact that it is difficult and expensive to obtain DNA sequences from archival tissues (Hajibabaei et al., 2005), new specimens are routinely collected for molecular phylogeographical studies. It is only natural that these additional samples will include new species.

There cannot be any doubt that the discovery of putatively new species based on DNA sequences creates many new opportunities in taxonomy. Once flagged as being genetically unusual, it is relatively easy for a taxonomist to focus on populations that are likely to represent new species. However, as I will discuss later in this chapter, the discovery of 'cryptic species' may currently actually create more problems than solutions.

DNA Sequences Help Assign Species Names to Tissues

Many biological tissues are routinely collected, but only a fraction can be identified to species. One problem is that species descriptions tend to focus on particular life history stages; for example, in insects it tends to be the adult male. The remaining material is often discarded because it appears unlikely that it will ever be useful given that it cannot be identified to species. Prominent examples include immature stages of arthropods, snails, fish and amphibians. The identification of all these can now be attempted using DNA sequences (Paquin and Hedin, 2004; Steinke et al., 2005; Vences et al., 2005a). Prior to the advent of DNA sequencing, the only avenue to getting determinations for immatures was rearing them to adulthood, which is impossible for many taxa and time consuming for the rest. Yet, these stages have always been interesting for many reasons. Systematists like to use immatures as a source of characters and ecologists have long known that for many taxa the immatures are ecologically more important than the adults. Ecologists also see considerable potential in using DNA sequences for identifying the gut content of native and introduced species (Kasper et al., 2004; Deagle et al., 2005); customs officers and epidemiologists like the ability of DNA sequences to predict future problems by providing names for immature stages (Armstrong and Ball, 2005), and medical doctors appreciate the quicker and cheaper diagnosis of fungal pathogens (Pryce et al., 2003; Summerbell et al., 2005). Here, DNA sequences are clearly an attractive 'new' identification tool that is potentially capable of identifying these hitherto unidentifiable tissues to species.

Similarly important is the use of DNA sequences in conservation biology and the food industries (Teletchea et al., 2005). All treaties that regulate the export and import of protected species will be useless unless custom officers can identify illegal material to species. However, few ports and airports have trained zoologists and botanists and even the trained staff may have problems distinguishing common and unprotected species from those that are rare and protected. The problem is exacerbated by the trade in animal and plant parts, which is growing quickly due to demand for traditional Chinese medicine. Many medicines contain ingredients from endangered species (e.g. tissues from tiger, rhinoceros, etc.) and molecular techniques have to be used to demonstrate that a particular product actually contains the contraband. Mislabelling is also a problem in the food industry and can be addressed with DNA sequences (Teletchea et al., 2005). However, despite all this promise, one should not expect that all tissues will yield useful DNA. The DNA obtained from many tissues is of very poor quantity and quality (e.g. treated with chemicals, cooked, etc.), the DNA sequence can be altered (Teletchea et al., 2005) and not all species have unique barcodes (Armstrong and Ball, 2005).

DNA Sequences Open the Door to Big Science Funding

The taxonomy crisis is to a large extent a funding crisis. Taxonomists have proper techniques for describing and identifying species, but taxonomy as a discipline lacks the necessary funding for accomplishing the task (Wilson, 2003; Wheeler et al., 2004). The main problem is that in its current form taxonomy cannot be easily automatized (Gaston and O'Neill, 2004). Worse still, not only does it need manpower but it also needs it in the form of highly trained personnel, usually holding PhD degrees. This has resulted in a funding gap between what is available and what is needed that by now is so substantial that only optimists could possibly think that it will be closed without getting funding from 'big science' budgets.

However, given the old-fashioned reputation of taxonomy, it appears very unlikely that this kind of funding will be available without a major image revamp. It is here that DNA sequences may help. Embracing molecular data may alleviate the image problem while at the same time providing important new data for addressing taxonomic problems. Of course, other innovations in taxonomy are arguably at least as important as DNA sequences. For example, new microscopic techniques (e.g. three-dimensional microscopy, remote microscopy) have improved the access and quality of the available information, interactive keys and automatization are revolutionizing species identification and the Internet is providing a convenient distribution platform (Godfray, 2002; Marshall, 2003; Gaston and O'Neill, 2004; Scoble, 2004; Wheeler, 2004; De Ley et al., 2005; Prendini, 2005; Will et al., 2005; Walter and Winterton, 2007). But it probably still remains easier to convince funding agency that molecular tools are the future of taxonomy; that is, 'big science' funding will probably not be available to taxonomy unless the field embraces molecular techniques (Hebert and Gregory, 2005).

DNA SEQUENCES IN TAXONOMY: CHALLENGES

The Nobel laureate Sydney Brenner once warned that scientists should not get 'infatuated with data for its own sake' and that if one does not define the problem, it will remain unclear 'what information is important'. He said that the idea that one can 'make a lot of measurements and something will come out of it' was 'rubbish' (Pennisi, 2003b, p. 1646). Yet, much of molecular taxonomy has shied away from clearly defining theoretically solid goals and much energy is invested in sequencing– an activity that Sydney Brenner considered 'so boring' that it should be done by prisoners as punishment: 'the more heinous the crime, the bigger the chromosome they would have to decipher' (Roberts, 2001). I have attended at least one talk by a molecular taxonomist during which he discussed new 'cryptic' species and listed developing a species concept as an important goal for future research. How did he know these sequences came from new species if he did not have a species concept? Worse still, some authors are not even interested in finding out whether units delimited based on DNA sequences represent species (Blaxter et al., 2005). Clearly, such approaches are not very useful and it is time to develop the necessary theoretical tools instead of focusing on the technological challenges (e.g. obtaining and processing sequences at a grand scale, developing 'tricoders' for sequencing, etc.). Therefore, I will here first discuss conceptual problems. For the sake of this discussion it is again necessary to distinguish between DNA barcoding and DNA taxonomy.

DNA Barcoding: Theoretical and Empirical Problems

In the molecular taxonomy literature DNA barcoding and DNA taxonomy are not clearly separated (but see Vogler and Monaghan, 2007), because the emphasis has been on presenting evidence from 'empirical' tests without clearly stating what was being tested. However, in its most extreme form, DNA barcoding is simply a 'new' identification technique for those species that have already been described by traditional taxonomists (Hebert et al., 2003a). Using DNA sequences for identifying these described species has several serious problems.

Barcoding Fails for Undescribed Species

It is generally thought that approximately 1.7 million species have been described during the first 250 years of taxonomy. However, for many hyperdiverse taxa, this is only 10–20 per cent of the estimated species diversity (Ødegaard, 2000). For example, only 50,000 of the estimated one million species of nematodes have names (Seberg, 2004) and recent revisions of invertebrates routinely double species counts (Ponder and Lunney, 1999). This indicates that collections contain at best identified material for 50 per cent of all species. Species richness estimation furthermore clearly indicates that many additional species remain uncollected and are thus not even represented in the natural history museums (Marshall, 2003; Meier and Dikow, 2004; Quicke, 2004; Seberg, 2004). Yet, all these undescribed species will have to be formally described before a reference sequence can be deposited in a DNA barcode database; that is, DNA barcoding is currently only of very limited value because it is applicable to only 10–20 per cent of all species. In particular, barcoding struggles for speciose taxa, although it is these that require most attention because specimens are particularly hard to identify.

What happens once a biologist tries to use DNA barcodes for identifying a specimen from an undescribed species? This depends on the identification tool that is used. Sophisticated tools will compare the query sequence from the specimen that needs identification with its best match in a barcode database and then decide whether the query is likely from an undescribed species or a species without known barcode. However, some identification tools will return an identification no matter how badly the query sequence from the specimen matches the sequences in the barcode database–that is, many queries will be misidentified (Meyer and Paulay, 2005; Meier et al., 2006). Given these problems with undescribed species, it does not come as a surprise that most of the current barcoding initiatives focus on taxa with very well-known faunas. The most extreme example is the current project on barcoding all bird species (Hebert and Gregory, 2005). Do we really need a new identification tools for birds given that it is relatively easy to use visual and auditory cues for species identification? Moreover, how are future birders going to obtain tissue samples for DNA barcoding?

Impossibility of Obtaining Barcodes for All Described Species

Hebert and Gregory (2005) recently announced that CBOL has already procured funding for 500,000 barcodes for the first 50,000 species; many biologists may conclude that it will only be a matter of time until all described species have been barcoded. However, it is relatively easy to obtain barcodes for the first 50,000 species, because they tend to be common and tissues are available from various collections. The real challenge is the 16.5 million samples for the next 1.65 million described species. Many specimens in museums are either unsuitable or too valuable for molecular work (Quicke, 2004; Will and Rubinoff, 2004). The extent of the problem becomes apparent when one realizes that approximately 40 per cent of all known beetle species have only been collected at a single locality (see Seberg, 2004). Many were killed using the DNA-damaging ethyl acetate and prior to pinning probably softened in a moist chamber for several days; i.e., much of their DNA will be degraded, which explains the low gene-amplification success for DNA extracted from pinned insect specimens (e.g. 31% for 'archival moths'; Hajibabaei et al., 2005). The material for other taxa is even less accessible. For example, in the past nematodes, mites and fish were mostly preserved in formaldehyde and/or slide mounted. Many nematode specimens have been lost and high-throughput sequencing of nematodes would destroy all new specimens, thus leaving no vouchers for future reference (De Ley et al., 2005). In addition, for many taxa only specimens identified during recent revisions are reliably determined and even obtaining DNA barcodes from specimens identified 25 years ago will yield many misidentified sequences (Meier and Dikow, 2004; Seberg, 2004). Given these problems, it is difficult to see how a complete barcode database could be created within a reasonable amount of time.

DNA Barcoding Is Dependent on Traditional Taxonomy

One of the main selling points of DNA barcoding has been the claim that it will be able to alleviate the crisis in taxonomy because it is 'faster' than traditional taxonomy. However, this claim cannot be maintained because only described species can be barcoded; that is, by definition DNA barcoding has to wait for traditional taxonomists to describe a species before it can be barcoded, so DNA barcoding will be even slower than traditional taxonomy. This is the main reason why proponents of barcoding are not purists and also want to use sequences for the discovery of new species (Hebert and Gregory, 2005).

Sources of Identification Errors: Barcodes from Misidentified Specimens

Ultimately, all barcodes are supposed to be deposited in a database (Hebert et al., 2003a). The obvious problem is that barcodes from misidentified specimens are also submitted. This has happened to Genbank (Ruedas et al., 2000; Harris, 2003; Hebert et al., 2003b; Vilgalys, 2003; Seberg, 2004) and will happen again to the barcode database because it is similar to Genbank in that many researchers will be allowed to submit. Even when good vouchering policies are implemented, it is unrealistic to believe that most vouchers will ever be re-identified by taxonomic experts and thus only the most glaring identification mistakes will be found (Ward et al., 2005). Furthermore, there are reasons for being particularly concerned about the proposals that DNA barcodes should be generated based on museum specimens and tissues from existing cryocollections, although the former contain many misidentified specimens (e.g. Meier and Dikow, 2004) and the latter frequently have poor vouchering policies. Judging from the problems in Genbank, misidentified barcodes will be a significant source of query misidentification.

Sources of Identification Errors: Shared Barcodes and Overlap between Intra- and Interspecific Variability

Biologists often assume that specimen identification based on DNA sequences is straightforward and essentially error free, but this known to be incorrect. Firstly, species can share barcodes. Meier et al. (2006) found for a Diptera data-set consisting of 1333 sequences for 449 species that even when two COI sequences were identical, there was a six per cent chance that they belonged to two different species; that is, shared barcodes were not as rare as had been argued before by Hebert and Gregory (2005). One prominent example is three of nine species in the *melanogaster* subgroup of *Drosophila* (Hurst and Jiggins, 2005), but similar cases are known from cichlids (Ferguson, 2002) and amphibians (Vences et al., 2005b). Shared barcodes may be more common than has been thought and they are also an underappreciated problem for associating life history stages and identifying biological tissues. It is often assumed that 'identical sequence' means 'same species', but this conclusion is not necessarily correct.

Shared species barcodes *sensu stricto* are even more common. Meier et al. (2006) argued that Hebert et al.'s DNA barcodes are really specimen barcodes because they do not summarize the sequence variability within a species. They thus proposed to construct species barcodes as the consensus sequence of all known sequences for a species. When they tested this approach in Diptera, they found that 21 per cent of all species for which multiple sequences were known shared the same consensus barcode. The ultimate source for this problem was the wide overlap between intraspecific and interspecific variability (0–17%) and, even when the largest five per cent of the intraspecific and the lowest five per cent of the interspecific values were excluded, the interval of overlap was still 2.31–3.34 per cent. Large overlap between intra- and interspecific variability is also known from other taxa, such as amphibians (Vences et al., 2005a,b), cowries (Meyer and Paulay, 2005) and even *Astraptes*, a taxon used to demonstrate the usefulness of DNA barcodes (Hebert et al., 2004a; Brower, 2006). This large intraspecific variability not only causes many problems in DNA-based identifications, but also questions the usefulness of the term 'DNA barcode' because it implies that each species has a fixed and invariant sequence (Moritz and Cicero, 2004). The well-established variability furthermore directly contradicts Hebert who, according to Pennisi (2003a, p. 1696), claimed that COI's 'sequence doesn't appear to vary among individuals of the same species'.

Uncertainty over Identification Success Rates

Identification success rates for DNA barcoding are a controversial topic in the literature. Several empirical tests have yielded rather low values in Diptera (<70%; Meier et al., 2006), Cypraeidae (cowries: 67–80%; Meyer and Paulay, 2005) and Cycadaceae (<70%; see chapter by D. Stevenson), while others have reported near perfect success rates (Hebert et al., 2003a, 2004b; Hogg and Hebert, 2004; Ball et al., 2005; Barrett and Hebert, 2005; Hebert and Gregory, 2005; Ward et al., 2005). What are the reasons for these discrepancies? As many authors have pointed out (Moritz and Cicero, 2004; Will and Rubinoff, 2004; Meyer and Paulay, 2005; Prendini, 2005; Meier et al., 2006), several 'successful tests' of barcoding were not sufficiently rigorous. For example, Hebert et al.'s (2003a, p. 314) most widely cited test used sequences from 'a single individual from each of the commonest lepidopteran species from a site near Guelph, Ontario, Canada'. Such a taxon sample will include few cases of closely related species. However, it is distinguishing closely related species that is difficult and a sample that is deficient in this regard will overestimate the success rate for DNA barcoding under realistic conditions (Sperling, 2003; Moritz and Cicero, 2004; Will and Rubinoff, 2004; Prendini, 2005).

The intraspecific variability was similarly inadequately sampled (Hebert et al., 2003a, 2004a; Ward et al., 2005). For most species only one specimen was sequenced and when additional specimens were included, they mostly came from the same population; that is, the intraspecific variability of COI was seriously underestimated. This inadequate sampling of intra- and interspecific variability led to an artificially small overlap between intra- and interspecific variability and thus to an overestimation of identification success. Fortunately, the recent barcoding literature indicates that newer studies are more careful about sampling the intraspecific variability by sequencing at least 10 specimens per species (Hajibabaei et al., 2005). However, the need for covering the entire geographical range of a species is apparently still not fully appreciated (Ward et al., 2005).

The high identification success rates for DNA barcodes in the literature are occasionally also due to the use of inadequate tree-based identification techniques. Proponents of DNA barcoding continue to consider query sequences successfully identified when they group most closely with their conspecific barcode, but this is obviously indefensible (Will and Rubinoff, 2004; Will et al., 2005). Imagine that a query sequence from an unidentified specimen clusters with a chimpanzee (*Pan*) barcode (Figure 7.1). Based on the query's position, one cannot decide whether it comes from *Homo sapiens* (Figure 7.1a) or another chimpanzee (Figure 7.1b); that is, forming a cluster on a tree is logically insufficient for identifying a sequence (Will and Rubinoff, 2004; Will et al., 2005). Instead, only those queries can be identified whose positions only allow for the assignment of a single species name to the query

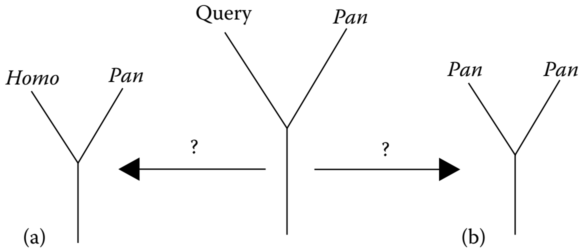


FIGURE 7.1 Grouping of a query with a barcode is insufficient for query identification. Query sequence can be from *Homo* (a) or *Pan* (b).

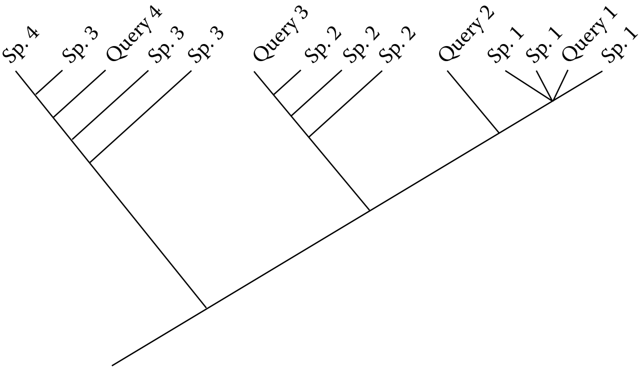


FIGURE 7.2 Tree-based query identification. Only queries 1, 3, and 4 can be identified according to revised tree-based identification criteria (see text for explanations).

sequence (Will and Rubinoff, 2004). These queries are found at least one node into a clade consisting of only sequences from the same species (e.g. Figure 7.2: 'Query 3'), are part of a polytomy formed by conspecific barcodes (e.g. Figure 7.2: 'Query 1') or are in a position that only allows for one parsimonious optimization of names (e.g. Figure 7.2: 'Query 4'). These are the rules that should govern tree-based sequence identification, but the obvious disadvantage is that they cannot yield correct identifications for those sequences that are in ambiguous positions (e.g. sistergroup to a species-specific clade; Figure 7.2: 'Query 2'). Note that given that many 'successful' tests of barcoding only included two sequences per species in analyses (e.g. Hebert et al., 2003a; Barrett and Hebert, 2005), the reported identification success rates should have been zero per cent.

An additional serious problem with Hebert et al.'s (2003a) original identification technique is that it implicitly assumes that species are monophyletic. However, the vast majority of species concepts neither require nor encourage 'monophyly' and the empirical evidence clearly demonstrates that species 'polyphyly' is rampant (Funk and Omland, 2003). The vast majority of described species were either described based on a concept stressing the importance of reproductive isolation (e.g. biological species concept, Hennigian species concept, cohesion species concept) or one based on unique combinations of character states (phylogenetic species concept *sensu* Wheeler and Platnick, 2000). Let us assume the following scenario. A subpopulation of a species acquires species status (e.g. through geographical separation), then the 'ancestral' species is 'paraphyletic' but all the aforementioned species concepts will nevertheless recognize it as a species (i.e. the sequences for the paraphyletic species will not cluster and thus be misidentified using Hebert's tree-based identification criteria). All this would not be a problem if these species concepts were rarely used, but they are the default concepts for the described species in the literature and DNA barcoding as a new identification tool for already described species is thus bound to frequently fail.

The empirical evidence supports this theoretical prediction. In a review of the phylogeography literature, Funk and Omland (2003) found that on gene trees 23 per cent of the 2319 surveyed species did not form monophyletic groups and that, in two-thirds of these cases, the polyphyly was even supported by bootstrap values above 70 per cent. For arthropods, the rate was even higher (26.5% of 702 species) and it is thus not surprising that Meier et al. (2006) also found a low identification success for DNA barcodes in Diptera. Overall, I find it difficult to reconcile the near-perfect identification successes reported in the barcoding literature with the data compiled by Funk and Omland (2003).

The *Homo-Pan* example demonstrated that species identification using trees has its problems and that identifying sequences entirely based on their position on an identification tree may not be possible. Instead, one generally also has to assess how well the nucleotide sequence of a query matches the sequence of its closest neighbor on the tree. Such approaches are thus increasingly adopted (see Ratnasingham and Hebert, 2007). Since such direct sequence comparison is also needed for tree-based identification, several authors have suggested that one may as well abandon trees and instead use direct sequence comparison for identifying specimens based on DNA sequences (Pozhitkov and Tautz, 2002; Blaxter et al., 2005; DeSalle et al., 2005; Steinke et al., 2005).

An example is Meier et al.'s (2006) 'best close match' criterion. First, they used pairwise distances to identify the best matching barcode(s) for a query. However, they only assigned the species name of the best matching barcode to the query if the barcode was sufficiently similar. In all other cases the query remained unidentified. This strategy required a threshold similarity value that defined how similar a barcode match had to be before the query could be identified. This value was estimated by obtaining a frequency distribution of all intraspecific pairwise distances and determining the threshold distance below which 95 per cent of all intraspecific distances were found. All queries without a sufficiently good match remained unidentified. Using this technique, Meier et al. (2006) found a 66 per cent identification success for their Diptera data-set (1333 sequences for 449 species). The main drawback of Meier et al.'s approach is that a given data-set may not be representative for the taxon under investigation. Furthermore, it also remains unclear why one should believe that there is a common threshold across all species. I thus sympathize with Summerbell et al.'s (2005, p. 1899) statement:

> Given that there can be no rules in biology of how quickly or slowly functional species may evolve, construction of a uniform barcode identification procedure may be regarded as a theoretical impossibility.

I am here emphasizing the problems that arise in developing universally applicable identification tools based on DNA sequences. This should not distract from the fact that DNA sequences are already very useful for less challenging tasks such as identifying species within a limited taxon sample or identifying a sequence to genus. Such applications are very common. For example, local faunas only contain a limited number of species with distinct barcodes, a customs official may not have to identify a piece of rhinoceros horn to species and only a relatively small number of fungi are of medical importance.

DNA Taxonomy: Theoretical and Empirical Problems

There are obvious problems with identifying described species using DNA barcoding. So, maybe the solution to the crisis in taxonomy is to use DNA sequences as the main tool for determining species boundaries and describing species. Several authors have proposed such a use of DNA sequences as a 'scaffold' of a future taxonomy. But the obvious questions are (1) how sequences can be clustered into DNA species and (2) whether these DNA sequence clusters capture species as conceptualized according to the various species concepts.

Theory: DNA Sequences and Species Concepts

There are several reasons why this topic is not popular with biologists in general and molecular taxonomists in particular (Sites and Marshall, 2003; Sanders et al., 2006). First, numerous species concepts have been described. Second, the literature on species concepts is vast and contentious. Third, all the debating has not yielded a consensus concept that is acceptable to all systematists. Thus, many molecular taxonomists are clearly avoiding this topic as much as they can. However, this is not a tenable position because the goal of taxonomy is to infer species boundaries and to describe species. Both are impossible without a species concept (Wheeler, 2005).

In the following paragraphs, I will investigate to what extent the techniques that have been used to delimit species based on DNA sequences can be matched to existing species concepts. For this purpose it is sufficient to consider three main classes of concepts. The first set is based on reproductive cohesion or isolation. The best known examples are the biological species concept (BSC; Mayr, 2000), the cohesion species concept (CSC; Templeton, 1989), the Hennigian species concept (HSC; Meier and Willmann, 2000), and the recognition species concept (RSC; Paterson, 1985). The second set is here represented by the phylogenetic species concept *sensu* (Wheeler and Platnick, 2000). It requires species to be assemblages of populations with unique combinations of characters. The third set here consists of the monophyletic species concept (MSC; Mishler and Theriot, 2000). In the literature it is also known as the phylogenetic species concept *sensu* Mishler and Theriot (2000). But in order to have only one "phylogenetic species concept" in the subsequent discussion, I will here call it MSC and use Mishler and Theriot's (2000, pp. 46–47) definition of the concept:

> A species is the least inclusive taxon recognized in a formal phylogenetic classification. As with all hierarchical levels of taxa in such a classification, organisms are grouped into species because of evidence of monophyly. Taxa are ranked as species rather than at some higher level because they are the smallest monophyletic groups deemed worthy of formal recognition, because of the amount of support for their monophyly and/or their importance in biological processes operating on the lineage in question.

The advantages and disadvantages of all these concepts have been discussed in great detail in Wheeler and Meier (2000) and several empirical studies have documented that the choice of species concept influences species numbers (Cracraft, 1992; Laamanen et al., 2003; Sanders et al., 2006).

Practice: How to Define Species Based on DNA Sequences

For the sake of this discussion it helps to envision a data-set that consists of one COI sequence for all animal and plant specimens on this planet. What we are looking for is a technique that groups all these sequences into discrete clusters that represent species according to a coherent species concept. Each sequence is only allowed to be a member of exactly one cluster and different biologists using the same clustering criterion (i.e. species concept) should find the same clusters given the same data.

Species Boundaries Based on Pairwise Distances. The first popular approach to defining species based on sequences uses pairwise distances or other measures of sequence similarity (Zhi et al., 1996; Chu et al., 1999; Tong et al., 2000; Blaxter, 2003; Chu et al., 2003; Sun et al., 2003; Tang et al., 2003; Oren, 2004; Shih et al., 2004; Monaghan et al., 2005; Smith et al., 2005). All specimens that yield sequences below a threshold value are considered conspecific. This approach requires a similarity threshold value. In animals the suggested value is often two to three per cent for COI (Hebert et al., 2003a, 2004b). However, assigning a value is increasingly untenable because two to three per cent would be considered low intraspecific variability in amphibians (Vences et al., 2005a) and snails (Steinke et al., 2005), while 2–3 per cent is too high a threshold for distinguishing species in Anthozoa (Hebert et al., 2003b). In addition, all these values are based on studies carried out on temperate species and are bound to change as more tropical taxa are included (Harris and Froufe, 2005). In summary, the choice of threshold was always somewhat questionable, but it is now considered downright arbitrary (DeSalle et al., 2005; Mallet et al., 2005) or it is even tautological when Hebert et al. (2004b) define it as 10 times the intraspecific variability. However, this arbitrariness is only a minor problem compared to three other issues:

1. Delimiting species based on pairwise distances cannot be justified based on any species concept that has ever been proposed. Standard population genetic theory predicts that once populations are separated, synonymous changes, mostly in third position of protein-encoding genes, they start accumulating.

Hence, populations separated for a relatively short time will have nearly identical COI sequences and populations separated for a long time will have large differences. Using a two to three per cent threshold value for species delimitation is thus equivalent to assuming that separated populations will automatically acquire species status after a fixed amount of time or, in other words, that species cannot be very young or very old (Ferguson, 2002). This assumption is at odds with 150 years of studying speciation and all species concepts. Consider those concepts based on reproductive isolation. Reproductive isolation is not acquired at an equal rate in all taxa (Ferguson, 2002), and thus delimiting species based on a pairwise distance thresholds of two to three per cent is bound to underestimate the species diversity in taxa with fast speciation and overestimate species diversity in taxa with slow changes in reproductive status. Pairwise distance threshold are also incompatible with the PSC of Wheeler and Platnick. Here, unique combinations of characters are required and they are again not logically connected to the amount of change. The MSC requires phylogenetic structure in the data and distances are non-specific about which nucleotide positions contribute to similarity and dissimilarity; that is, sequences within the same 2 per cent cluster may be found in very different clades on a phylogenetic tree.

2. The second objection to threshold-based species is closely related to the first. Supporters of such species have justified their technique by pointing out that, on average, described congeneric species differ by two to three per cent; that is, this threshold can be used to separate species. This approach would only be meaningful if speciation were a clockwise process. Then we would not expect to ever find 'old' species with large intraspecific variability and all species would be separated from their next relative by approximately the same COI distance. However, speciation is not clockwise and we thus find a wide range of distances.

   Thus, the real challenge in taxonomy is to properly recognize species boundaries between sister species and to recognize polymorphic species. Examples for the former are several sympatric species of *Sapromyza* (Lauxaniidae: Diptera) and *Drosophila* with identical COI sequences (Pestano et al., 2003; Hurst and Jiggins, 2005) and cases like the Sumatran and Bornean orangutan subspecies, which differ by six per cent for COI and nevertheless freely hybridize when given the chance (Xu and Arnason, 1996; Muir et al., 1998). Average congeneric interspecific distance values do not help in all of these cases. It is also important to point out that the relatively low average values reported for different animal taxa are somewhat dubious because molecular taxonomists tend to explain away higher than 'normal' intraspecific distances as evidence for cryptic species; that is, artificially low intraspecific distances are maintained by considering large distances taxonomic error (Hebert et al., 2004b; Hogg and Hebert, 2004; Ball et al., 2005; Barrett and Hebert, 2005; Smith et al., 2005).

3. A largely neglected problem of distance-based species delimitations is that they are not even logically consistent. Three sequences can have two pairwise distances conforming to and one exceeding a given threshold

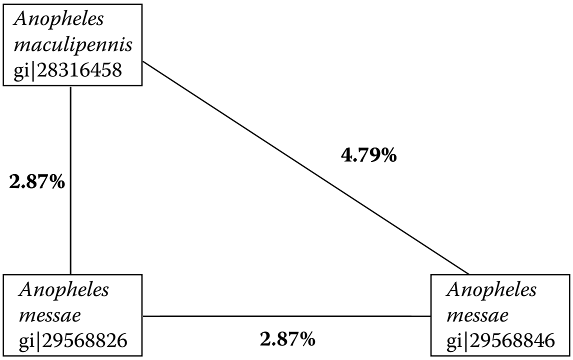

Anopheles maculipennis gi|28316458
2.87%
4.79%
Anopheles messae gi|29568826
2.87%
Anopheles messae gi|29568846

FIGURE 7.3 Pair-wise distances for three *Anopheles* sequences. Two sequences have one distance complying and one distance violating the 3 per cent threshold.

(Figure 7.3). Under this circumstance, it remains unclear whether all three sequences should be included in the same DNA species given that one pairwise distance is violating the threshold. Such threshold violations are more than just a theoretical problem. Meier et al. (2006) found them to be very common in their Diptera data-set regardless of what pairwise distance value was used as a threshold. For example, 11 per cent of the 123 DNA clusters base don a 3 per cent threshold had pairwise distances in excess of the threshold. The largest observed distance in a 3 per cent cluster was 4.8 per cent. Meier et al. (2006) also found that the majority of the three per cent clusters contradicted traditionally recognized species with some species having sequences in six different DNA clusters. One could argue that the problem is not the DNA clusters, but the traditionally defined species. However, here it is important to remember what it means that two COI sequences differ by 4 per cent. In many cases, the two sequences will still code for exactly the same amino acid sequence; for a biologist should border on the bizarre to argue that two specimens should be reassigned to two different species because they have an identical COI protein that happens to be coded for by a slightly different DNA sequence.

Species Boundaries Based on Gene Trees. Using trees for determining species boundaries is getting increasingly popular. However, it is problematic for sexually reproducing organisms, as was first pointed out by Hennig (1966). Within such species the relationships between populations and individuals are usually not hierarchical. Forcing nuclear DNA sequences from these individuals and populations into a tree is thus at best questionable (Davis and Nixon, 1992). Mitochondrial data appear at first to be less problematic because the maternal inheritance of mitochondria ensures that the data were produced via a hierarchical process. However, the relationships on gene trees for mitochondrial genes do at best reflect the maternal side of the population history and a proper taxonomic classification should go beyond reflecting female relationships. This is particularly important for those species where females and males have very different dispersal behaviors (Moore, 1995; Moritz and

Cicero, 2004; Hendrixson and Bond, 2005). Apart from these concerns, we need to investigate how species boundaries could be drawn based on trees and whether the resulting units are compatible with the different species concepts.

Various techniques have been proposed for determining species boundaries on trees. Most have the following underlying goals (Wiens and Penkrot, 2002). Species should be monophyletic on the tree and the clades denoted as species should be reasonably well supported in order to be deemed worthy of being recognized as a species. Several problems with these criteria have been pointed out. One is the old 'metaspecies' problem (Mishler and Brandon, 1987; Willmann and Meier, 2000)–that is, cases where one clade is well supported as a species, but the support for the rest of the tree is so poor that it cannot be used to assign the remaining sequences to monophyletic species. Under these circumstances, some specimens are not members of any monophyletic species and are thus provisionally assigned to a 'metaspecies'. Additional problems arise when branch length (Monaghan et al., 2005) or branch support measures are used to determine which branch is worthy of being considered a species. First, the choice of threshold is again largely arbitrary (Willmann and Meier, 2000) and difficult given that branch support values often have a continuous distribution across a tree. For example, the branch lengths and bootstrap values observed in Meier et al.'s (2006) Diptera data-set do not allow for an unambiguous division of the trees into species-rank clades. Furthermore, many well-supported clades are found within currently accepted species (see Figures 7.4 and 7.5). Second, more data in the form of additional genes will undoubtedly uncover additional well-supported branches and new species will have to be described. For example, sequencing complete mitochondrial genomes for many individuals in a population may eventually yield a well-supported tree that resolves the relationships between all mitochondria within that population. But why should a group of individuals sharing a haplotype be recognized as a species?

Are species boundaries drawn based on gene trees compatible with mainstream species concepts? As discussed previously they are compatible with the MSC, a concept that I do not think is well reasoned given that the 'deemed worthy' criterion is subjective and the MSC is forcing non-hierarchical data onto hierarchical trees. Species boundaries determined based on gene trees are incompatible with all other species concepts. This has been explicitly expressed with regard to the PSC (Nixon and Wheeler, 1990; Davis and Nixon, 1992), but also applies to all species concepts based on reproductive isolation and cohesion. This is important because most users of taxonomy (i.e. biologists in various disciplines) expect species to be reproductively isolated units. Yet, there is mounting evidence that reproductively isolated units generally comprise multiple species once the MSC is applied (Sanders et al., 2006). For example, based on a review of genetic data for 250 vertebrate species, Avise and Walker (1999) estimated that the number of species would probably double if tree-based species concepts were applied. Prominent examples would include *Homo sapiens* and *Pongo,* which both have well-supported intraspecific branches (Zhi et al., 1996). Yet, the populations are clearly capable of interbreeding and all supporters of species concepts based on reproductive isolation would not split *Pongo pygmaeus* or *Homo sapiens* into several species (Muir et al., 1998). Ballard et al. (2002) similarly

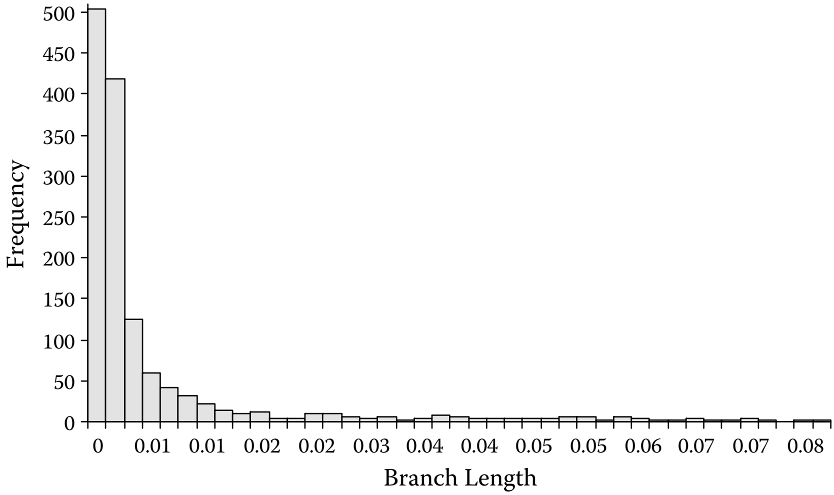


FIGURE 7.4 Combined distribution of branch lengths on neighbour-joining trees for analyses of COI sequences from 17 genera of Diptera (*Aedes* 32seq./14spp.; *Anastrepha* 45seq./15spp.; *Anopheles* 188seq./11spp.; *Asphondylia* 42seq./3spp; *Bactrocera* 27seq./2spp.; *Calliphora* 15seq./2spp.; *Chiastocheta* 50seq./9spp.; *Chironomus* 43seq./34spp.; *Chrysomya* 16seq./9spp; *Culicoides* 86seq./6spp.; *Liriomyza* 19seq./4spp.; *Lucilia* 50seq./7spp.; *Paragus* 12seq./7spp.; *Phytomyza* 19seq./14spp.; *Sapromyza* 32seq./11spp.; *Scathophaga* 35seq./15spp.; *Drosophila s*ubgenus *Drosophila* 93seq./33spp., subgenus *Sophophora* 111seq./47spp.).

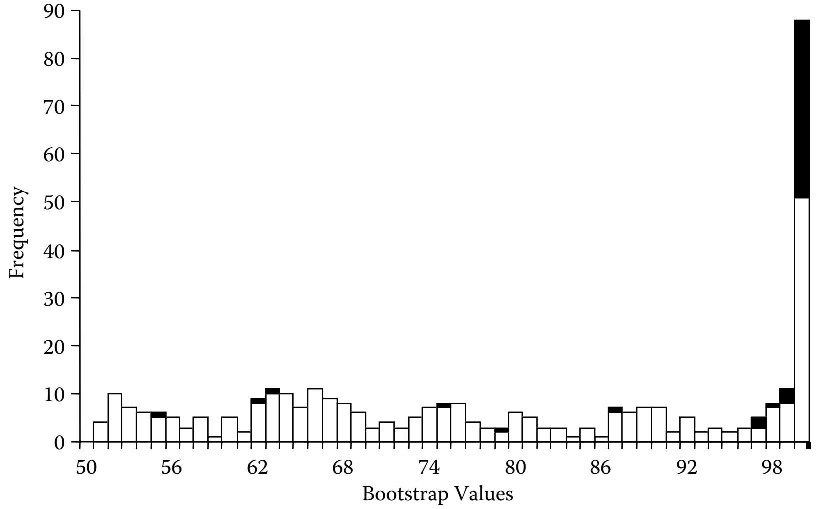


FIGURE 7.5 Combined distribution of branch support (b) on gene trees for congeneric sequences (see Figure 7.2); black bar = bootstrap support for currently recognized species.

demonstrated that *Drosophila simulans* consisted of three mitochondrial lineages that would be considered 'species' according to the MSC, although all three are capable of interbreeding. Overall, we thus have to conclude that there is widespread conflict between species as reproductive units and tree-based species.

I am thus not impressed with the techniques that have been proposed for using DNA sequences as a scaffold for creating a new taxonomy. Determining species boundaries based on overall similarity is bound to fail because this technique is not logically consistent unless all distances between sets of three sequences are equidistant. Determining species boundaries based on gene trees has a host of other problems and the units that will be defined as species have little correspondence to what most biologists consider species. Some authors have proposed that DNA sequences 'can be used as a "triage" tool for sorting new collections into units' (Schindel and Miller, 2005, p. 17) and that subsequent proper taxonomic work would resolve the species status. But this approach depends on having proper clustering techniques and I am here arguing that they remain elusive. Furthermore, all species sharing barcodes will be overlooked, although these taxa are arguably particularly interesting from an evolutionary point of view.

DNA Sequences in Taxonomy: Other Challenges

Which Genes Should Be Used?

The list of genes that have been used for taxonomic purposes is long and contains mitochondrial, chloroplast and nuclear genes. Mitochondrial genes are the clear favorite in animals, because mitochondrial DNA is easily amplified and evolves quickly enough to yield sequence differences between closely related species. Among the mitochondrial genes, CBOL has been arguing for COI; however, vertebrate systematists, especially, remain unconvinced and have evidence for a better performance of 16S rDNA (Vences et al., 2005a,b). In any case, sequences from the mitochondrial genome are not without their drawbacks. Because of maternal inheritance, interbreeding cannot be directly inferred from mitochondrial gene sequences and in species with unequal dispersal ability between males and females, mitochondrial gene trees are not an appropriate estimate for the entire species (Moore, 1995; Moritz and Cicero, 2004; Hendrixson and Bond, 2005). Nuclear copies of mitochondrial genes, lineage sorting, introgression and other phenomena can produce spurious signal (Bensasson et al., 2001; Ballard et al., 2002; Funk and Omland, 2003; Moritz and Cicero, 2004; Thalmann et al., 2004; Sanders et al., 2006) and recent research has uncovered alarmingly frequent sweeps of mitochondrial haplotypes through infections with endosymbionts such as *Wolbachia* or *Cardinium*. These sweeps are so common that Hurst and Jiggins (2005, p. 1526) concluded that 'mtDNA is inappropriate as a sole marker in studies of the recent history of arthropods and, potentially, other invertebrates'.

Given all these problems with mitochondrial data, it does not come as a surprise that conflict between mitochondrial and other data is common (Burbrink et al., 2000; Lee, 2000; Puorto et al., 2001; Ballard et al., 2002; Bensch et al., 2006; Sanders et al., 2006). This conflict and many theoretical problems render it unwise to only use mitochondrial sequences in DNA barcoding or use them as the scaffold of a DNA taxonomy. The obvious solution is adding sequences from the nuclear genome (Moore, 1995; Sperling, 2003; Hurst and Jiggins, 2005). Especially, nuclear ribosomal genes (Blaxter et al., 2005; Markmann and Tautz, 2005) and ITS have been discussed (Summerbell et al., 2005). However, the former tend to be rather conservative and the latter has problems with multiple different copies. Furthermore, for most nuclear genes the amplification success is lower (Monaghan et al., 2005). I would nevertheless argue that eukaryote systematists can learn from the experiences of prokaryote taxonomists who have a long history of using DNA sequences for taxonomic purposes. Here, the ad hoc committee for the re-evaluation of the species definition in bacteriology now recommends that at least five housekeeping genes be sequenced for new species (Stackebrandt et al., 2002).

Bioinformatic Challenges

DNA barcoding and DNA taxonomy are already generating thousands of sequences every year. These sequences need to be processed before they can be used in research. The first processing step that deals with the chromatograms produced by the sequencer will not be discussed here, although even mass processing techniques still require some manual editing (Hajibabaei et al., 2005). Once one consensus sequence per specimen has been obtained, the sequences have to be compared. All initial work appears to have been based on explicit multiple sequence alignments. However, there is now a tendency in the literature to use more approximate methods such as Blast (Kasper et al., 2004; Hajibabaei et al., 2005; Vences et al., 2005a,b). It appears to me that this is very dangerous. Blast may be fast, but it is also known to be imprecise (Koski and Golding, 2001; Nilsson et al., 2004) and approximate techniques are likely to yield incorrect results given the relatively small sequence differences between closely related species.

Another problem is that tree-based identification is currently based on NJ trees, although NJ trees have numerous problems (Will and Rubinoff, 2004; Prendini, 2005). For example, data ambiguity is difficult to detect, because most NJ algorithms only generate a single tree even if other trees have the same fit to the data ('tie trees'; see Takezaki, 1998). Tree choice then becomes dependent on taxon entry order, which is hardly acceptable for a technique used in science (Backeljau et al., 1996). Simulation studies have furthermore revealed that tie trees are particularly prominent on trees with short internal branches (Takezaki, 1998) and these can be expected to be common on identification trees, given that they usually have multiple sequences from the same or closely related species.

Scale and Cost

I doubt that the scale of the proposed DNA barcoding project has been appropriately appreciated. Hajibabaei et al. (2005) argue that 10 specimens should be sequenced for an estimated 10 million species of eukaryotes. This would generate a barcode database that is twice the size of today's Genbank (Hajibabaei et al., 2005). Seberg (2004) has argued that handling each sequence would take at least 5 minutes. Sequence handling alone will thus require 4166 years of labor, assuming 250 eight-hour workdays per year. Now add the time needed for recording specimen label data, voucher processing, DNA extraction, and data analysis and it appears unlikely that DNA barcoding and DNA taxonomy will be faster than traditional taxonomy. Yet, I am still excluding the time needed for specimen acquisition. CBOL now considers 10 million species its target number, because it is now targeting not only the described but also the undescribed species. Ten specimens for each species would have to be collected and sorted to morphospecies, because otherwise it would be impossible to obtain exactly 10 sequences for each species.

Of course, these 10 specimens should come from throughout the range of the species, so it will be necessary to compare morphospecies across different collecting sites. Such projects have been tried at a smaller scale and the logistics have been found to be very taxing. In contrast to traditional taxonomy, which can rely on museum specimens, most specimens for the DNA barcoding project will have to be freshly collected because museum specimens have too low a sequencing success for the mass production of barcodes (Hajibabaei et al., 2005). An additional problem is that the majority of the species live in the tropics and many tropical countries lack the infrastructure for large-scale sequencing. Furthermore, the cost for molecular research in such countries is often much higher than in temperate countries and obtaining research and collecting permits alone will require many years.

Given all these problems, it is very difficult to estimate the cost for the project, but all figures mentioned in the literature are seriously misleading because they exclude labor (Will et al., 2005). Here are some numbers: Stoeckle (2003) is the cheapest and promises amplification and sequencing for USD 1 per sequence; Hebert and Gregory (2005) estimate USD 2; Ball et al. (2005) USD 5; and Tautz et al. (2003) 5 Euro. More careful estimates have been prepared by Pryce et al. (2003), who used DNA sequences for diagnosing fungal diseases. They calculated a cost of AUD 10.12 (USD 7.50) for reagents and AUD 6.54 for labor (USD 4.90). The total cost for 100 million specimens and sequences would thus be approximately 500 million dollars for reagents, but this is only a fraction of the real cost, which will have to include travel, collecting, labor, etc.

Sequences for Cryptic Species

DNA sequences obtained for putatively ‘cryptic species’ hidden within described species are quickly developing into a first-rate taxonomic problem (Hebert et al., 2003a; Prendini, 2005). First, there is no standardized policy with regard to how these sequences should be submitted to GenBank. I have seen submissions under the name of the described species without indication that the sequence may come from a new species (e.g. Frost et al., 1998), with an indication of uncertainty using an ‘aff.’ (e.g. Feder et al., 2003) or ‘cf’ (e.g. Sharpe et al., 2000), or under the genus name with the addition of a ‘sp.’ (with number: e.g. Smith et al., 2005; without numbering: e.g. Hebert et al., 2004a). Second, most of the cryptic species are never described (Ebach and Holdrege, 2005). Opinions are evenly divided on whether DNA sequences are sufficient for this purpose (Murray and Stackebrandt, 1995; Bond and Sierwald, 2003; Tautz et al., 2003; Lee, 2004; Dayrat, 2005; Hebert and Gregory, 2005), but the literature clearly indicates that the vast majority of molecular taxonomists are unwilling to formally assign names to cryptic species, which are thus quickly becoming ghosts in the literature. Since every biologist is allowed to submit a sequence under, for example, *Drosophila* sp. 1, 'sp. 1' will be an ambiguous 'name' and we are back to where we were when Linnaeus introduced taxonomic nomenclature in order to avoid having different biologists using the same name for different species. Taxonomists with classical training are more likely to avoid this pitfall and use sequences within their species descriptions (Bond and Sierwald, 2003; Bond, 2004). Given the inconsistent and improper treatment of DNA sequences for putatively cryptic species, I have to question Hebert and Gregory's suggestion (2005, p. 853) that DNA sequences 'aid taxonomic investigations'. Taxonomists do not suffer from a paucity of species that need description and systematists are certainly not waiting for more cryptic species.

One of the worst offenders with regard to submitting unidentified sequences is CBOL member institutions. As of February 2006, 5426 sequences had been submitted to GenBank that contain the keyword 'barcode'. Of these, 2459 (45%) were only identified to genus. What is urgently needed is a standardized policy with regard to cryptic species. Given the evidence that many haplotype-based lineages do not correspond to what most biologists consider species, I would hope that the term 'cryptic species' disappears from the literature and is replaced with ESU (evolutionary significant unit; e.g. Paetkau, 1999), which is a better description of what the data support. Alternatively, one could again learn from prokaryote taxonomy and introduce a 'Candidatus' status that has long been available for putatively new prokaryote species that lack the required data for a formal description (Oren, 2004; Murray and Stackebrandt, 1995).

Vouchering

Given the widespread problem with "cryptic" species, it is all the more important that proper vouchering policies are put in place. It was heartening to see that the latest literature on DNA barcoding clearly recognizes the need for proper vouchering and that a significant amount of funding will be dedicated for storing vouchers in museums (Hebert and Gregory, 2005; Janzen et al., 2005). However, not all molecular taxonomists are paying sufficient attention to vouchering and, especially in soft-bodied invertebrates, voucher-eating molecular techniques are still widely used (Floyd et al., 2002; Blaxter et al., 2004, 2005), although better alternatives are available (De Ley et al., 2005).

DNA SEQUENCES IN TAXONOMY: OUTLOOK

When the Human Genome project was proposed almost 20 years ago, critics called it 'absurd', 'dangerous' and 'impossible' and one of the questions asked was, "Who would want the complete sequence data … even if the project's starry-eyed proponents could by some miracle pull it off?' (Roberts, 2001). Yet, the Human Genome project went ahead and yielded a wealth of knowledge. Could it be that the critics of DNA barcoding and DNA taxonomy are as short-sighted as the critics of the Human Genome project? A more complete review of the history of this project would reveal that many scientists argued that the project was too expensive and that it would be more efficient to spend the same funding on many smaller projects. But reality teaches that those who fund ‘big science’ are not interested in small projects. Yet, taxonomy needs ‘big science’ funding and the question is thus which kind of bold taxonomy vision will be eligible for big science funding.

Is it going to be the more traditional approaches represented by the All Species Foundation or NSF’s PEET (Smith, 2005) or PBIs grants (Planetary Biodiversity Inventory; Wheeler, 2004) or is it going to be DNA barcoding and/or DNA taxonomy? To me the former are more scientifically sound, but they have been unsuccessfully arguing for big science funding for several years. It appears to me that proposals that emphasize DNA sequences are more likely to be successful because their proponents are more willing to promise too much in too little time based on ‘modern’ techniques. This is the kind of proposal that is popular with organizations that fund big science. The Human Genome project took the same approach and only toned down its claims after the funding was approved. Furthermore, some of the more questionable claims were silently dropped and in the end the results were still very useful.

I would hope that the same will happen to DNA barcoding and that a compromise can be reached. It appears to me that giving DNA sequences prominence is the price that taxonomy has to pay in order to get back onto its feet. In any case, the positions of supporters and critics may not be as far apart as it seems. For example, I doubt that even the staunchest critic of DNA barcoding will argue that there is something inherently ‘evil’ about having millions of COI sequences that can be used for taxonomic purposes, and CBOL members have already paid lip service to the importance of funding collection-based research. The logical compromise would be an integrative taxonomy based on morphological and molecular data, which would satisfy all parties and furthermore make scientific sense. Let us thus learn from the physicists who have been able to attract big science funding by giving the whole community a share in the project, appearing united to the outside world and keeping the critical discourse contained within the community. To reach this goal it is important to overcome personal animosities and it is here that CBOL does not have the best track record because some members have been dividing the world into friends and enemies and tried to sideline critics who are not likely to turn into supporters of the initiative as long as ‘extravagant’ (Ebach and Holdrege, 2005) and clearly unsupported claims are constantly repeated (e.g. near perfect identification success based on inappropriate tree-based techniques). I believe that it is time to tone down the rhetoric and embrace a more integrative approach to taxonomy.

Advantages of Integrative Taxonomy: Data Quality and Quantity

There is no doubt that taxonomy can profit from the additional information provided by DNA sequence data (Dayrat, 2005; Page et al., 2005). DNA sequences can potentially also be important for identifying biological tissues and associating different life history stages. However, DNA sequences alone have numerous shortcomings that have been discussed at length in this chapter. It is thus preferable to combine sequences from different genes and genomes with morphological information (Will et al., 2005; Roe and Sperling, 2007). By not only gathering information on COI but also inspecting morphology, a larger proportion of the genome is sampled at the phenotypic level (Will et al., 2005). This integrative approach yields benefits for molecular and traditional taxonomists alike. Taxonomic problems can only be detected when information from multiple sources is available; that is, the weakness of any one-dimensional approach is avoided (Bond, 2004; Paquin and Hedin, 2004, Will et al., 2005; Roe and Sperling, 2007). I would thus argue that the real frontiers in taxonomy are not DNA taxonomy or DNA barcoding, but finding ways to combine the DNA sequence data with morphology, systematically studying how common conflict is between the partitions, and finding appropriate ways to resolve the conflict (Wiens and Penkrot, 2002; Sites and Marshall, 2003; Page et al., 2005). Thus, more pilot studies like those of Ballard et al. (2002), Wiens and Penkrot (2002), Sanders et al. (2006) and Roe and Sperling (2007), which address these issues, are urgently needed.

Advantages of Integrative Taxonomy: History and Data Accessibility

The natural history museums of the world already contain millions of specimens. Many belong to undescribed species, but are not suitable for sequencing. This is clearly the result of Hajibabaei et al.'s (2005) systematic study of amplification success in 'archival' moths. If we were to use DNA sequences as the main scaffold of a future taxonomy, we may as well discard the museum specimens because the 'scaffold' will never be available for them. It is thus important that also in the future species description include morphology, or in other words, that the old and the new taxonomy remain connected. However, this connection can only be maintained through an integrative approach that combines morphology and molecules. The morphological information will be all the more important because not all parties interested in identifying specimens have access to sequencing facilities (Dunn, 2003; Prendini, 2005). For example, we should not forget that the vast majority of species live in tropical countries and that many have neither high-throughput sequencing facilities nor the funding for such research; that is, requiring sequences for taxonomy would effectively kill taxonomic activity in many countries (Will et al., 2005).

ACKNOWLEDGEMENTS

RM would like to thank the Systematics Association and organizer of the 'New Taxonomy' Symposium, Dr. Quentin Wheeler, for the invitation to participate. RM also gratefully acknowledges manuscript comments by Dr Peter Ng and financial support by the Systematics Association and the Academic Research Fund grant R-154-000-256-112 from the Ministry of Education, Singapore.

REFERENCES

Armstrong, K.F. and Ball, S.L. (2005) DNA barcodes for biosecurity: Invasive species identification. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1813–1823.

Avise, J.C. (2000) *Phylogeography: The History and Formation of Species,* Harvard University Press, Cambridge, MA.

Avise, J.C. and Walker, D. (1999) Species realities and numbers in sexual vertebrates: Perspectives from an asexually transmitted genome. *Proceedings of the National Academy of Sciences of the United States of America,* 96: 992–995.

Backeljau, T., De Bruyn, L., De Wolf, H., Jordaens, K., Van Dongen, S. and Winnepenninckx, B. (1996) Multiple UPGMA and neighbour-joining trees and the performance of some computer packages. *Molecular Biology and Evolution,* 13: 309–313.

Ball, S.L., Hebert, P.D.N., Burian, S.K. and Webb, J.M. (2005) Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society,* 24, 508–524.

Ballard, J.W.O., Chernoff, B. and James, A.C. (2002) Divergence of mitochondrial DNA is not corroborated by nuclear DNA, morphology, or behavior in *Drosophila simulans. Evolution,* 56: 527–545.

Barrett, R.D.H. and Hebert, P.D.N. (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology,* 83: 481–491.

Bensasson, D., Zhang, D.-X., Hartl, D.L. and Hewitt, G.M. (2001) Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology and Evolution,* 16: 314–321.

Bensch, S., Irwin, D.E., Irwin, J.H., Kvist, L. and Akesson, S. 2006. Conflicting patterns of mitochondrial and nuclear DNA diversity in *Phylloscopus* warblers. *Molecular Ecology,* 15: 161–171.

Bequaert, J.C. (1953) The Hippoboscidae or louse-flies (Diptera) of mammals and birds. Part 1. Structure, physiology and natural history. *Entomologica Americana (New Series),* 32: 1–209.

Bequaert, J.C. (1954) The Hippoboscidae or louse-flies (Diptera) of mammals and birds. Part II. Taxonomy, evolution and revision of American genera and species. *Entomologica Americana (New Series),* 34: 1–232.

Bickford, D., Lohman, D., Sodhi, N.S., Ng, P.K.L., Meier, R., Winker, K., Ingram, K. and Das, I. (2007) Cryptic species: A new window on diversity and conservation. *Trends in Ecology and Evolution,* 22: 148–155.

Birstein, V.J., Doukakis, P., Sorkin, B. and Desalle, R. (1998) Population aggregation analysis of three caviar-producing species of sturgeons and implications for the species identification of black caviar. *Conservation Biology,* 12: 766–775.

Blaxter, M. (2003) Counting angels with DNA. *Nature,* 421: 122–124.

Blaxter, M., Elsworth, B. and Daub, J. (2004) DNA taxonomy of a neglected animal phylum: An unexpected diversity of tardigrades. *Proceedings of the Royal Society of London Series B-Biological Sciences,* 271: S189–S192.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R. and Abebe, E. (2005) Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1935–1943.

Bond, J.E. (2004) Systematics of the Californian euctenizine spider genus *Apomastus* (Araneae: Mygalomorphae: Cyrtaucheniidae): The relationship between molecular and morphological taxonomy. *Invertebrate Systematics,* 18: 361–376.

Bond, J.E. and Sierwald, P. (2003) Molecular taxonomy of the *Anadenobolus excisus* (Diplopoda: Spirobolida: Rhinocricidae) species-group on the Caribbean island of Jamaica. *Invertebrate Systematics,* 17: 515–528.

Brower, A.V.Z. (2006) Problems with DNA barcodes for species delimitation: 'Ten species' of *Astraptes fulgerator* reassessed (Lepidoptera: Hesperiidae). *Systematics and Biodiversity,* 4: 127–132.

Burbrink, F.T., Lawson, R. and Slowinski, J.B. (2000) Mitochondrial DNA phylogeography of the polytypic North American rat snake (*Elaphe obsoleta*): A critique of the subspecies concept. *Evolution,* 54; 2107–2118.

Caterino, M.S., Cho, S. and Sperling, F.A.H. (2000) The current state of insect molecular systematics: A thriving tower of Babel. *Annual Review of Entomology,* 45: 1–54.

Chu, K.H., Ho, H.Y., Li, C.P. and Chan, T.Y. (2003) Molecular phylogenetics of the mitten crab species in *Eriocheir, sensu lato* (Brachyura: Grapsidae). *Journal of Crustacean Biology,* 23: 738–746.

Chu, K.H., Tong, J. and Chan, T.Y. (1999) Mitochondrial cytochrome oxidase I sequence divergence in some Chinese species of *Charybdis* (Crustacea: Decapoda: Portunidae). *Biochemical Systematics and Ecology,* 27: 461–468.

Cracraft, J. (1992) The species of the birds of paradise Paradisaeidae. Applying the phylogenetic species concept to a complex pattern of diversification. *Cladistics,* 8: 1–43.

Davis, J.I. and Nixon, K.C. (1992) Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology,* 41: 421–435.

Dayrat, B. (2005) Towards integrative taxonomy. *Biological Journal of the Linnean Society,* 85: 407–415.

Deagle, B.E., Jarman, S.N., Pemberton, D. and Gales, N.J. (2005) Genetic screening for prey in the gut contents from a giant squid (*Architeuthis* sp.). *Journal of Heredity,* 96: 417–423.

De Ley, P., De Ley, I.T., Morris, K., Abebe, E., Mundo-Ocampo, M., Yoder, M., Heras, J., Waumann, D., Rocha-Olivares, A., Burr, A.H.J., Baldwin, J.G. and Thomas, W.K. (2005) An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1945–1958.

DeSalle, R., Egan, M.G. and Siddall M. (2005) The unholy trinity: Taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society (Biological Sciences),* 360: 1905–1916.

Dunn, C.P. (2003) Keeping taxonomy based in morphology. *Trends in Ecology & Evolution,* 18: 270–271.

Ebach, M.C. and Holdrege, C. (2005) DNA barcoding is no substitute for taxonomy. *Nature,* 434: 697.

Feder, J.L., Berlocher, S.H., Roethele, J.B., Dambroski, H., Smith, J.J., Perry, W.L., Gavrilovic, V., Filchak, K.E., Rull, J. and Aluja, M. (2003) Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis. *Proceedings of the National Academy of Sciences of the United States of America,* 100: 10314–10319.

Ferguson, J.W.H. (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society,* 75: 509–516.

Floyd, R., Abebe, E., Papert, A. and Blaxter, M. (2002) Molecular barcodes for soil nematode identification. *Molecular Ecology,* 11: 839–850.

Frost, D.R., Crafts, H.M., Fitzgerald, L.A. and Titus, T.A. (1998) Geographic variation, species recognition, and molecular evolution of cytochrome oxidase I in the *Tropidurus spinulosus* complex (Iguania: Tropiduridae). *Copeia,* 1998: 839–851.

Funk, D.J. and Omland, K.E. (2003) Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology and Systematics,* 34: 397–423.

Gaston, K.J. and O'Neill, M.A. (2004) Automated species identification: Why not? *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences,* 359: 655–667.

Gibbs, M.J., Armstrong, J.S. and Gibbs, A.J. (2005) Individual sequences in large sets of gene sequences may be distinguished efficiently by combinations of shared subsequences. *BMC Bioinformatics,* 6: 90.

Godfray, H.C.J. (2002) Challenges for taxonomy. *Nature,* 417: 17–19.

Godfray, H.C.J. and Knapp, S. (2004) Taxonomy for the twenty-first century–Introduction. *Philosophical Transactions of the Royal Society of London B Biological Sciences,* 359: 559–569.

Hajibabaei, M., DeWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M. and Hebert, P.D.N. (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1959–1967.

Harris, D.J. (2003) Can you bank on GenBank? *Trends in Ecology & Evolution,* 18: 317–319.

Harris, D.J. and Froufe, E. (2005) Taxonomic inflation: Species concept or historical geopolitical bias? *Trends in Ecology & Evolution,* 20: 6–7.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society (Biological Sciences Series),* 270: 313–321.

Hebert, P.D.N. and Gregory, T.R. (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology,* 54: 852–859.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. and Hallwachs, W. (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America,* 101: 14812–14817.

Hebert, P.D.N., Stoeckle, M., Zemlak, T.S. and Francis, C.M. (2004b) Identification of birds through DNA barcodes. *PLoS Biology,* 2: 1657–1663.

Hebert, P.D.N., Ratnasingham, S. and deWaard, J.R. (2003b) Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society Biological Sciences Series B,* 270: S96–S99.

Hendrixson, B.E. and Bond, J.E. (2005) Testing species boundaries in the *Antrodiaetus unicolor* complex (Araneae: Mygalomorphae: Antrodiaetidae): 'Paraphyly' and cryptic diversity. *Molecular Phylogenetics and Evolution,* 36: 405–416.

Hennig, W. (1966) *Phylogenetic Systematics,* University of Illinois Press, Urbana, 280 pp.

Hill, D.S. (1962a) Revision of the British *Ornithomyia* (Diptera: Hippoboscidae). *Proceedings of the Royal Entomological Society of London (Series B),* 31: 11–18.

Hill, D.S. (1962b) A study of the distribution and host preference of three species of *Ornithomyia* (Diptera: Hippoboscidae) in the British Isles. *Proceedings of the Royal Entomological Society of London (Series A),* 37: 37–48.

Hill, D.S. (1963) The life history of the British species of *Ornithomya* (Diptera: Hippoboscidae). *Transactions of the Royal Entomological Society of London,* 115: 391–407.

Hill, D.S., Hackmann, W. and Lyneborg, L. (1964) The genus *Ornithomya* (Diptera: Hippoboscidae) in Fennoscandia. Denmark and Iceland, *Notulae Entomologicae,* 44: 33–52.

Hogg, I.D. and Hebert, P.D.N. (2004) Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology,* 82: 749–754.

Hurst, G.D.D. and Jiggins, F.M. (2005) Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: The effects of inherited symbionts. *Proceedings of the Royal Society (Series B),* 272: 1525–1534.

Hutson, A.M. (1984) Keds, flat-flies and bat-flies. *Handbooks for the Identification of British Insects,* 10: 1–40.

Janzen, D.H., Hajibabaei, M., Burns, J.M., Hallwachs, W., Remigio, E. and Hebert, P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society (Series B),* 360: 1835–1845.

Johnsen, P. (1948) Notes on the Danish louse-flies (Diptera: Hippoboscidae). *Entomologiske Meddelelser,* 25: 278–298.

Kasper, M.L., Reeson, A.F., Cooper, S.J.B., Perry, K.D. and Austin, A.D. (2004) Assessment of prey overlap between a native (*Polistes humilis*) and an introduced (*Vespula germanica*) social wasp using morphology and phylogenetic analyses of 16S rDNA. *Molecular Ecology,* 13: 2037–2048.

Koski, L.B. and Golding, G.B. (2001) The closest BLAST hit is often not the nearest neighbour. *Journal of Molecular Evolution,* 52: 540–542.

Krzywinski, J. and Besansky, N.J. (2003) Molecular systematics of *Anopheles*: From subgenera to subpopulations. *Annual Review of Entomology,* 48: 111–139.

Laamanen, T.R., Petersen, F.T. and Meier, R. (2003) Kelp flies and species concepts: The case of *Coelopa frigida* (Fabricius, 1805) and *C. nebularum* Aldrich, 1929 (Diptera: Coelopidae). *Journal of Zoological Systematics and Evolutionary Research,* 41: 127–136.

Lee, C.E. (2000) Global phylogeography of a cryptic copepod species complex and reproductive isolation between genetically proximate 'populations'. *Evolution,* 54: 2014–2027.

Lee, M.S.Y. (2004) The molecularization of taxonomy. *Invertebrate Systematics,* 18: 1–6.

Mallet, J., Isaac, N.J.B. and Mace, G.M. (2005) Response to Harris and Froufe, and Knapp et al.: Taxonomic inflation. *Trends in Ecology & Evolution,* 20: 8–9.

Markmann, M. and Tautz, D. (2005) Reverse taxonomy: An approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philosophical Transactions of the Royal Society of London (Series B),* 360: 1917–1924.

Marshall, S. (2003) The real costs of insect identification. *Newsl. Biol. Surv. Can. (Terrestrial Arthropods), Opin. Page,* 22: 1–4.

Mayr, E. (2000) The biological species concept. In *Species Concepts and Phylogenetic Theory: A Debate* (eds Q.D. Wheeler and R. Meier), Columbia University Press, New York, pp. 17–29.

Meier, R. and Dikow, T. (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable and complete specimen data. *Conservation Biology,* 18: 478–488.

Meier, R., Kwong, S., Vaidya, G. and Ng, P.K.L. (2006) DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology,* 55: 715–728.

Meier, R. and Willmann, R. (2000) The Hennigian species concept. In *Species Concepts and Phylogenetic Theory: A Debate* (eds Q.D. Wheeler and R. Meier), Columbia University Press, New York, pp. 30–43.

Memon, N., Meier, R., Mannan, A. and Su Feng-Yi, K. (2006) On the use of DNA sequences for determining the species limits of a polymorphic new species in the stink bug genus *Halys* (Heteroptera: Pentatomidae) from Pakistan. *Systematic Entomology,* 31: 703–710.

Meyer, C.P. and Paulay, G. (2005) DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology,* 3: 2229–2238.

Mishler, B. and Theriot, E. (2000) The phylogenetic species concept *sensu* Mishler and Theriot: Monophyly, apomorphy, and phylogenetic species concepts. In *Species Concepts and Phylogenetic Theory: A Debate* (eds Q.D. Wheelerand R. Meier), Columbia University Press, New York, pp. 44–54.

Mishler, B.D. and Brandon, R.N. (1987) Individuality, pluralism, and the phylogenetic species concept. *Biology and Philosophy,* 2: 397–414.

Monaghan, M.T., Balke, M., Gregory, T.R. and Vogler, A.P. (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1925–1933.

Moore, W.S. (1995) Inferring phylogenies from mtDNA variation: Mitochondrial-gene trees versus nuclear-gene trees. *Evolution,* 49: 718–726.

Moritz, C. and Cicero, C. (2004) DNA barcoding: Promise and pitfalls. *PLoS Biology,* 2: 1529–1531.

Muir, C.C., Galdikas, B.M.F., Beckenbach, A.T. and Arnason, U. (1998) Is there sufficient evidence to elevate the orangutan of Borneo and Sumatra to separate species? (and reply). *Journal of Molecular Evolution,* 46: 378–381.

Murray, R. and Stackebrandt, E. (1995) Taxonomic note: Implementation of the provisional status Candidatus for incompletely described procaryotes. *International Journal of Systematic Bacteriology,* 45: 186–187.

Nilsson, R.H., Larsson, K.H. and Ursing, B.M. (2004) Galaxie–CGI scripts for sequence identification through automated phylogenetic analysis. *Bioinformatics,* 20: 1447–1452.

Nixon, K.C. and Wheeler, Q.D. (1990) An amplification of the phylogenetic species concept. *Cladistics,* 6: 211–224.

Ødegaard, F. (2000) How many species of arthropods? Erwin's estimate revised. *Biological Journal of the Linnean Society,* 71: 583–597.

Oren, A. (2004) Prokaryote diversity and taxonomy: Current status and future challenges. *Philosophical Transactions of the Royal Society of London (Series B),* 359: 623–638.

Paetkau, D. (1999) Using genetics to identify intraspecific conservation units: A critique of current methods. *Conservation Biology,* 13: 1507–1509.

Page, T.J., Choy, S.C. and Hughes, J.M. (2005) The taxonomic feedback loop: Symbiosis of morphology and molecules. *Biology Letters,* 1: 139–142.

Palumbi, S.R. and Cipriano, F. (1998) Species identification using genetic tools: The value of nuclear and mitochondrial gene sequences in whale conservation. *Journal of Heredity,* 89: 459–464.

Paquin, P. and Hedin, M. (2004) The power and perils of 'molecular taxonomy': A case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. *Molecular Ecology,* 13: 3239–3255.

Paterson, H.E.H. (1985) The recognition species concept of species. In *Transvaal Museum Monograph* (ed E. Vrba), Transvaal Museum, Pretoria, pp. 21–29.

Pennisi, E. 2003a. Modernizing the tree of life. *Science,* 300: 1692–1697.

Pennisi, E. 2003b. Systems biology: Tracing life's circuitry. *Science,* 302: 1646–1649.

Pestano, J., Brown, R.P., Suarez, N.M. and Baez, M. (2003) Diversification of sympatric *Sapromyza* (Diptera: Lauxaniidae) from Madeira: Six morphological species but only four mtDNA lineages. *Molecular Phylogenetics and Evolution,* 27: 422–428.

Petersen, F.T., Damagaard, J. and Meier, R. (2007a) How many DNA sequences are needed for solving a taxonomic problem? The case of two parapetic species of house flies (Diptera: Hippoboscidae: Ornithomya Latreille, 1802). *Arthropod Systematics & Phylogeny*, 65: 111–117.

Petersen, F.T., Meier, R., Kutty, S.N. and Wiegmann, B.M. (2007b) The phylogeny and evolution of host choice in the Hippoboscoidea (Diptera) as reconstructed using four molecular markers. *Molecular Phylogenetics and Evolution,* 45: 111–122.

Ponder, W. and Lunney, D. (1999) *The Other 99%–The Conservation and Biodiversity of Invertebrates,* Royal Zoological Society of New South Wales, Sydney, 462 pp.

Pozhitkov, A.E. and Tautz, D. (2002) An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification. *BMC Bioinformatics,* 3: 1–7.

Prendini, L. (2005) Comment on 'Identifying spiders through DNA barcodes'. *Canadian Journal of Zoology,* 83: 498–504.

Pryce, T.M., Palladino, S., Kay, I.D. and Coombs, G.W. (2003) Rapid identification of fungi by sequencing the ITSI and ITS2 regions using an automated capillary electrophoresis system. *Medical Mycology,* 41: 369–381.

Puorto, G., da Graca Salomao, M., Theakston, R.D.G., Thorpe, R.S., Warrell, D.A. and Wuster, W. (2001) Combining mitochondrial DNA sequences and morphological data to infer species boundaries: Phylogeography of lanceheaded pitvipers in the Brazilian Atlantic forest, and the status of *Bothrops pradoi* (Squamata: Serpentes: Viperidae). *Journal of Evolutionary Biology,* 14: 527–538.

Quicke, D.L.J. (2004) The world of DNA barcoding and morphology–Collision or synergism and what of the future? *The Systematist–Newsletter of the Systematics Association,* 8-12: 90. (www.systass.org/newsletter/)

Ratnasingham, S. and Hebert, P.D.N. (2007) BOLD: The barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes,* 7: 355–364.

Roberts, L. (2001) Controversial from the start. *Science,* 291: 1182–1188.

Roe, A.D. and Sperling, F.A.H. (2007) Population structure and species boundary delimitation of cryptic *Dioryctria* moths: An integrative approach. *Molecular Ecology,* 16: 3617–3633.

Ruedas, L.A., Salazar-Bravo, J., Dragoo, J.W. and Yates, T.L. (2000) The importance of being earnest: What, if anything, constitutes a 'specimen examined'? *Molecular Phylogenetics and Evolution,* 17: 129–132.

Sanders, K.L., Malhotra, A. and Thorpe, R.S. (2006) Combining molecular, morphological and ecological data to infer species boundaries in a cryptic tropical pit viper. *Biological Journal of the Linnean Society,* 87: 343–364.

Schindel, D.E. and Miller, S.E. (2005) DNA barcoding useful for taxonomists. *Nature,* 435: 17.

Scoble, M.J. (2004) Unitary or unified taxonomy? *Philosophical Transactions of the Royal Society of London (Series B),* 359: 699–710.

Scotland, R., Hughes, C., Donovan, B. and Wortley, A. (2003) The Big Machine and the much-maligned taxonomist. *Systematics and Biodiversity,* 1: 139–143.

Seberg, O. (2004) The future of systematics: Assembling the Tree of Life. *The Systematist–Newsletter of the Systematics Association,* 2-8: 23. (www.systass.org/newsletter)

Sharpe, R.G., Harbach, R.E. and Butlin, R.K. (2000) Molecular variation and phylogeny of members of the *Minimus* group of *Anopheles* subgenus *Cellia* (Diptera: Culicidae). *Systematic Entomology,* 25: 263–272.

Shih, H.-T., Ng, P.K.L. and Chang, H.-W. (2004) The systematics of the genus *Geothelphusa* (Crustacea, Decapoda, Brachyura, Potamidae) from southern Taiwan: A molecular appraisal. *Zoological Studies,* 43: 561–570.

Sites, J.W., Jr. and Marshall, J.C. (2003) Delimiting species: A Renaissance issue in systematic biology. *Trends in Ecology & Evolution,* 18: 462–470.

Smith, M.A., Fisher, B.L. and Hebert, P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: The ants of Madagascar. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1825–1834.

Smith, V.S. (2005) DNA barcoding: Perspectives from a 'Partnerships for Enhancing Expertise in Taxonomy' (PEET) debate. *Systematic Biology,* 54: 841–844.

Sodhi, N.S., Koh, L.P., Brook, B.W. and Ng, P.K.L. (2004) Southeast Asian biodiversity: An impending disaster. *Trends in Ecology & Evolution,* 19: 654–660.

Sperling, F. (2003) DNA barcoding. Deus ex machina. *Newsl. Biol. Surv. Can. (Terrestrial Arthropods) Opin. Page,* 22: 50–53.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P., Kampfer, P., Maiden, M., Nesme, X., Rossello-Mora, R., Swings, J., Truper, H.G., Vauterin, L., Ward, A.C. and Whitman, W.B. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology,* 52: 1043–1047.

Steinke, D., Vences, M., Salzburger, W. and Meyer, A. (2005) TaxI: A software tool for DNA barcoding using distance methods. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1975–1980.

Stoeckle, M. (2003) Taxonomy, DNA and the Bar Code of Life. *BioScience,* 53: 796–797.

Summerbell, R.C., Levesque, C.A., Seifert, K.A., Bovers, M., Fell, J.W., Diaz, M.R., Boekhout, T., de Hoog, G.S., Stalpers, J. and Crous, P.W. (2005) Microcoding: The second step in DNA barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1897–1903.

Sun, H.Y., Zhou, K. and Yang, X.J. (2003) Phylogenetic relationships of the mitten crabs inferred from mitochondrial 16S rDNA partial sequences (Crustacean, Decapoda). *Acta Zoologica Sinica,* 49: 592–599.

Takezaki, N. (1998) Tie trees generated by distance methods of phylogenetic reconstruction. *Molecular Biology and Evolution,* 15: 727–737.

Tang, B., Zhou, K., Song, D., Yang, G. and Dai, A. (2003) Molecular systematics of the Asian mitten crabs, genus *Eriocheir* (Crustacea: Brachyura). *Molecular Phylogenetics and Evolution,* 29: 309–316.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2003) A plea for DNA taxonomy. *Trends in Ecology & Evolution,* 18: 70–74.

Teletchea, F., Maudet, C. and Hanni, C. (2005) Food and forensic molecular identification: Update and challenges. *Trends in Biotechnology,* 23: 359–366.

Templeton, A.R. (1989) The meaning of species and speciation: A genetic perspective. In *Speciation and Its Consequences* (eds D. Otte and J. Endler), Sinauer, Sunderland, MA, pp. 3–27.

Thalmann, O., Hebler, J., Poinar, H.N., Pääbo, S. and Vigilant, L. (2004) Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Molecular Ecology,* 13: 321–335.

Theodor, O. and Oldroyd, H. (1964) *Hippoboscidae,* E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, Germany, 126 pp.

Tong, J.G., Chan, T.Y. and Chu, K.H. (2000) A preliminary phylogenetic analysis of *Metapenaeopsis* (Decapoda: Penaeidae) based on mitochondrial DNA sequences of selected species from the Indo West Pacific. *Journal of Crustacean Biology,* 20: 541–549.

Vences, M., Thomas, M., Bonett, R.M. and Vieites, D.R. (2005a) Deciphering amphibian diversity through DNA barcoding: Chances and challenges. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1859–1868.

Vences, M., Thomas, M., Van der Meijden, A., Chiari, Y. and Vieites, D.R. (2005b) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology,* 2: 5.

Vilgalys, R. (2003) Taxonomic misidentification in public DNA databases. *New Phytologist,* 160: 4–5.

Vogler, A.P. and Monaghan, M.T. (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research,* 45: 1–10.

Walter, D.E. and Winterton, S. (2007) Keys and the crisis in taxonomy: Extinction or reinvention? *Annual Review of Entomology,* 52: 193–208.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. and Hebert, P.D.N. (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360: 1847–1857.

Wheeler, Q.D. (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society of London B Biological Sciences,* 359: 571–583.

Wheeler, Q.D. (2005) Losing the plot: DNA 'barcodes' and taxonomy. *Cladistics,* 21: 405–407.

Wheeler, Q.D. and Meier, R., eds (2000) *Species Concepts and Phylogenetic Theory: A Debate,* Columbia University Press, New York, 256 pp.

Wheeler, Q.D., and Platnick, N.I. (2000) The phylogenetic species concept *sensu* Wheeler and Platnick. In *Species Concepts and Phylogenetic Theory: A Debate* (eds Q.D. Wheeler and R. Meier), Columbia University Press, New York, pp. 55–69.

Wheeler, Q.D., Raven, P.H. and Wilson, E.O. (2004) Taxonomy: Impediment or expedient? *Science,* 303: 285.

Wiens, J.J. and Penkrot, T.A. (2002) Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology,* 51: 69–91.

Will, K.W. and Rubinoff, D. (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics,* 20: 47–55.

Will, K.W., Mishler, B.D. and Wheeler, Q.D. (2005) The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology,* 54: 844–851.

Willmann, R. and Meier, R. (2000) A critique from the Hennigian species perspective. In *Species Concepts and Phylogenetic Theory: A Debate* (eds Q.D. Wheeler and R. Meier), Columbia University Press, New York, pp. 101–118.

Wilson, E.O. (2003) The encyclopedia of life. *Trends in Ecology & Evolution,* 18: 77–80.

Xu, X. and Arnason, U. (1996) The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. *Journal of Molecular Evolution,* 43: 431–437.

Zhi, L., Karesh, W.B., Janczewski, D.N., Frazier-Taylor, H., Sajuthi, D., Gombek, F., Andau, M., Martenson, J.S. and O'Brien, S.J. (1996) Genomic differentiation among natural populations of orangutan (*Pongo pygmaeus*). *Current Biology,* 6: 1326–1336.

8 Animal Names for All

ICZN, ZooBank and the New Taxonomy

Andrew Polaszek, Richard Pyle and Doug Yanega

CONTENTS

INTRODUCTION

Linnaeus revolutionized animal nomenclature with his binominal system introduced in the 10th edition of *Systema Naturae* (Linnaeus, 1758). A quarter of a millennium later we are still without a complete and authoritative catalogue of all scientific names of animals. Fortunately, current advances in information technology at last make it possible to provide this resource so fundamental to the zoological sciences. The organisation responsible for maintaining the standards and quality control required for the production of this catalogue–an index to the encyclopaedia of animal life–is the International Commission on Zoological Nomenclature, ICZN.

ICZN, 'the Commission', was founded in 1895 as a result of an awareness among zoologists of an increasing degree of chaos and controversy in the scientific naming of animals, and the lack of any universally accepted guidelines to ensure stability in zoological nomenclature. The Commission presently comprises 28 commissioners from 20 countries, periodically elected by their fellow zoologists, normally at International Congresses of Zoology, which take place every 4 years. The composition of the Commission aims to reflect broad zoological and geopolitical coverage, and the Commission is governed by its constitution and bylaws.

Much of the Commission's routine work is managed by the ICZN Secretariat, based at various times in Washington D.C. and London, where it is presently housed in the Natural History Museum. The Commission, including the Secretariat, receives neither governmental nor international funding, and its finances are managed by ITZN ('the Trust'), which is itself a registered UK charity. Funding is obtained largely through journal subscriptions and donations–a business model that will change over the next few years as the Commission's products and services become increasingly web based and open access.

ICZN's mission is 'to achieve stability and sense in the scientific naming of animals', and the Commission endeavours to do this by assisting the zoological community through the generation and dissemination of information on the correct use of the scientific names of animals. During this process the Commission acts as both advisor and arbiter. Outputs include publication of the *International Code of Zoological Nomenclature*, and the *Bulletin of Zoological Nomenclature*, the latter containing applications to, and rulings by, the Commission. ICZN aims to distribute this information as widely as possible, working towards the provision of a free service. Both the Convention on Biological Diversity (CBD) and the International Union of Biological Sciences (IUBS) provide the Commission with clear mandates for its activities–the former via Decisions IV/1/8 and VIII/3 that include, respectively, the following aims:

> Institutions, supported by Parties and international donors, should coordinate their efforts to establish and maintain effective mechanisms for the stable naming of biological taxa [and] [a] widely accessible checklist of known species.
>
> **(CBD, 2005)**

The IUBS has consistently endorsed the Commission's activities by adopting each edition of the *International Code of Zoological Nomenclature*. Since June 2005 ICZN has been an associate participant of the Global Biodiversity Information Facility (GBIF), and thereby also a member of the GBIF governing board. ICZN has been a member of the Taxonomic Databases Working Group (TDWG) since October 2006.

The present roles of ICZN are largely twofold: to periodically revise and publish the *International Code of Zoological Nomenclature* and to consider and rule on specific cases of nomenclatural uncertainty in zoology. The code is now in its fourth edition (1999), the first dating from 1960. It is structured as a series of articles, sections and subsections allocated to several chapters. It is a highly complex document, running to over 150 pages, and phrased generally in language that can be described as quasi-legalistic. In late 2004 a searchable, fully cross-referenced, web version of the code was placed on the ICZN website (www.iczn.org/iczn/index.jsp).

The Commission has produced the *Bulletin of Zoological Nomenclature* since 1943, and its content consists largely of cases, comments and opinions. Where a problem is discovered concerning the naming of animals, applicants present an argument, called a *case,* which is published in the *Bulletin* and usually requires a decision by the Commission. Counter arguments or support (*comments*) are published subsequently and the Commission then votes on the application. The final result of this process is that the Commission's decision (assuming a two-thirds majority vote), called an *opinion,* is then published in the *Bulletin.* Many, if not most, cases involve the routine conservation of animal names that would otherwise fall as junior homonyms, or objective or subjective synonyms, following the *principle of priority* that in essence states that older names take precedence over younger ones. However, articles relating to frequency of usage are a very important aspect of the fourth edition of the code and complicate the issue. Despite the undoubted intention of upholding the principles of stability embedded in the code, many cases require an extended process involving considerable numbers of hours put in by the authors, ICZN Secretariat and the Commissioners, while the outcome may affect only a handful of interested individuals. Methods for streamlining this process are therefore currently being developed by the Secretariat, and these will eventually result in online submission forms for the simpler cases.

One of the by-products of the Commission's rulings, or opinions, is that certain names are thereby placed on the *Official Lists and Indexes of Names in Zoology.* Essentially, the conserved names go on the *Lists*, while the rejected names go on the *Indexes*. These *Lists* and *Indexes* cover family, genus and species group names, and there are additional lists of approved or rejected 'works' (i.e. publications). Compilations of the *Lists* and *Indexes* have been published in hard copy and are also available as pdf files on the ICZN website at www.iczn.org/Official_Lists_Indexes.htm. The information contained in these lists is extremely important in terms of data quality, and code compliance.

Code compliance determines whether or not the scientific name of an animal can or cannot be used correctly in any context–whether it is 'available' in ICZN terminology. Adherence to the code is an example of unity in the zoological sciences, and efforts must be maintained to preserve this situation while we undergo our taxonomic renaissance and set the stage for the new taxonomy. Code compliance confers confidence in the correct usage of animal names, and this has far-reaching implications, particularly whenever these names are used in a legalistic context. Thus, wherever identification, authentication, verification and quality control are required to a 'legal' standard, the authority of the code provides an endorsement of the quality of the animal names or nomenclatural acts involved. For example, where these names are being used in the context of international trade in animals or animal products, quarantine, conservation (e.g. CITES, red data lists), biodiversity studies, medical and veterinary research or biostratigraphy, an additional level of authority–that conferred by code compliance–may be required.

The fundamental principles in the code are rather few, but are complicated by individual histories and, more recently, by exceptions based on usage. Criteria of name availability are complex and depend initially on what actually constitutes publication under the code, and subsequently on a whole series of parameters that have to be satisfied. The present code does not regard a situation where the primary means of publication is electronic as fulfilling the criteria of availability (ICZN, 1999: Article 9.8) and this is currently causing problems for online journals that publish animal taxonomy. As stated before, the *principle of priority* is central to the code, and it in turn affects synonymy and homonymy. Also central to the code is the *principle of coordination,* whereby within the family, genus, or species group, a name established for a taxon at any rank in the group is simultaneously established with the same author and date for taxa based on the same name-bearing type at the other ranks. The type specimen concept is also a fundamental tenet of the code, but as with the criteria for what constitutes publication, some ambiguity is present. For example, see the arguments presented by Wakeham-Dawson et al. (2002), Polaszek et al. (2005a) and Dubois and Nemésio (2007). These two issues of electronic publication and typification are currently at the core of the debate in animal nomenclature.

ICZN–FUTURE ROLE

The renaissance that taxonomy is currently experiencing results from a combination of factors including, particularly, ease of DNA/RNA sequencing, imaging technology and large-scale relational databasing capacity. These factors, coupled with the Internet for communication, are poised to create an arena for a revolution in organismal taxonomy and systematics. Taxonomy is a discipline that requires web technology to express itself at its best. Pictures paint thousands of words, but are expensive to publish in hard copy–the web completely overcomes this constraint. Graphics software allows for the production and dissemination of images of a quality hitherto unimaginable. Databasing of taxonomic information in a variety of interaccessible formats is now possible on an unprecedented scale. Furthermore, the actual process of describing new taxa will be greatly facilitated by web tools, whereby character/taxon matrices are used to produce taxon descriptions, differential diagnoses and identification tools, and taxonomic mark-up schemas, such as TaxonX and TaXMLit (see http://research.amnh.org/informatics/taxlit/schemas).

An era of digital taxonomy as envisaged by, among others, Godfray (2002a,b) and Wilson (2003), with an explicit agenda to accelerate the description of the world's rapidly dwindling biodiversity at a rate orders of magnitude higher than has been the case for the last 250 years, will require an effective, reliable and unambiguous code that can facilitate this process. Furthermore, the treatment of nomenclatural problems via the traditional cases, comments and opinions, as described previously, will need to be greatly streamlined and shifted entirely to the web, except for paper archiving of opinions. The most pressing requirements for facilitating web taxonomy are thus an improved user-friendly code, a web-based system for dealing with nomenclatural problems to replace the cases and opinions and, most importantly, a register of all animal names.

ZOOBANK: A UNIVERSAL REGISTER FOR ANIMAL NAMES

It is remarkable that in 2007 we still do not have a single authoritative list of all animal generic names, let alone a species list. Despite the large number of commendable initiatives to produce databases of animal scientific names, we are still without such a basic tool. Mandatory registration of new animal names, in parallel with retrospective registration of existing names, would provide such a list, which by its very nature would be always complete and up to date. Registration is not a new idea; it has existed for bacteria since 1980, was briefly introduced for plants from 1998–1999 and has been in place as a voluntary system for fungi with MycoBank since 2004. Registration of animal names was proposed by Zoological Record in 2003 (Thorne, 2003), but was never established. The idea was recently revived, as 'ZooBank', by Polaszek et al. (2005b,c) and has since gained approval and support from the zoological community. The following is a summary argument for the establishment of ZooBank as an open-access web register for animal taxa and nomenclatural acts.

Descriptions of new animal species and associated nomenclatural acts are currently published in thousands of specialized journals, monographs and CDs. Many of these publications are difficult to obtain, and the taxonomic acts contained in them are sometimes obscure. Establishing a mandatory register of these names and acts, as is proposed with the creation of ZooBank, will greatly increase the 'visibility' and, consequently, the availability of these data. ZooBank will eventually make all nomenclatural data in zoology freely available and will also provide an alerting service targeting taxa of interest to particular user groups. Making registration of new names a mandatory ICZN requirement for their availability, coupled with retrospective registration of existing names, will ensure the completeness of ZooBank.

Having code compliance built into the registration process provides an opportunity to introduce unprecedented stability into zoological nomenclature. The ZooBank interface will automatically check for code compliance and thus prevent new homonymy, stabilize spellings, fix stems and solve many problems concerning gender. As well as increased stability, the ZooBank register will provide an opportunity for increased quality control in animal nomenclature. ZooBank will enable the tracking of names and hence facilitate the correction of many problems prior to publication and name availability. This will in turn address and correct many areas of ambiguity such as the necessity for type specimens, type depositories, offensive names, auctioning or otherwise selling names and/or type specimens, and other ethical issues.

With animal taxonomy moving away from the traditional media of printed journals or monographs towards the Internet, the mandatory register of names assumes a central role. We have found that the consensus view among zoologists is that without a mandatory register for animal names and nomenclatural acts, web taxonomy would rapidly become unmanageable. ZooBank will thus facilitate 'true' web taxonomy–that is, taxonomy that exists only on the Internet. However, in order for web taxonomy to become a reality, an effective and fair peer-review system still needs to be developed, and several possible models for this process are discussed later.

ZooBank could also be a depository of taxon descriptions, which would thereby be universally available. Including original descriptions as a mandatory part of registering new taxa would be difficult to achieve, partly for reasons of current copyright laws. ZooBank will, however, provide a voluntary field for original descriptions (e.g. in pdf format), with no limit on numbers of illustrations. Having these descriptions freely available, with links to the original papers, will greatly benefit both authors and publishers. A number of prominent publishers of animal taxonomy have already agreed to make such information available to ZooBank. There are several other initiatives currently active with the long-term aim of developing comprehensive databases of animal names. These include ITIS (Integrated Taxonomic Information System) and Species 2000 (together forming the Catalogue of Life initiative), uBIO, Zoological Record/ION and ECAT (see previous discussion). ZooBank's uniqueness as an (eventually) mandatory animal name register distinguishes it from these projects, while having considerable complementarity with them.

Registering names with ZooBank will be via an online registration form available to authors or third parties, which is currently in its developmental phase. Standard fields for taxonomic data will be included, as well as additional fields dealing with code compliance, type depositories, gender and stem. As well as these mandatory fields, optional fields would be provided for type locality details, and, as discussed before, the description and figures. Registration will be based on GUIDs/DOIs (globally unique identifiers/digital object identifiers), specifically utilizing the life science identifier (LSID) system. Development of LSIDs for ZooBank will be undertaken in close cooperation with GBIF and TDWG. Prepublication registration of new animal names and nomenclatural acts has several parallels with the accession number system for gene sequences in GenBank.

Also, as with GenBank, editors and publishers will require authors to register new descriptions and nomenclatural acts with ZooBank. Publishers will also be strongly encouraged to allow the inclusion in ZooBank of descriptive/nomenclatural sections of published work. A holding period, with a maximum of 2 years, will be necessary during the prepublication phase during which as yet unavailable names will not be openly accessible. Code-compliance checks will be built into the registration process, and registration will remain free to all users. In order to prevent casual, routine or even unscrupulous registration of names that are unlikely to ever be published, prepublication registration will require the completion of all mandatory fields and publication within 2 years. The ZooBank ‘prototype’, consisting of *Zoological Record*’s Index of Organism Names data via a ZooBank portal (www.zoobank.org) was made available online in August 2006.

Botanists attempted to introduce a registration system for plants at the time of the Tokyo Botanical Congress (1994) but this was not ratified at the following St. Louis Congress (1999), although a voluntary system ran for a number of years. The reasons why plant and fungal taxonomists failed to adopt mandatory registration are partly that there are far fewer plant names to deal with (about 1/10 of animal names), and also that there is already a very effective universal checklist of plant names in the form of the International Plant Name Index (IPNI–http://www.ipni.org/index.html). For bacteria, a mandatory registration system has been in place since 1980. Bacterial names are considered to be available only if published in the *International Journal of Systematic and Evolutionary Microbiology* (formerly *International Journal of Systematic Bacteriology*–http://ijs.sgmjournals.org).

A dedicated discussion list on ZooBank has been established at http://list.afriherp.org/mailman/listinfo/zoobank-list, and from the contributions it is possible to gauge levels of acceptance by the zoological community. While the majority of comments to date have been constructive and supportive, some contributors have suggested that mandatory registration of organismal names is authoritarian and/or imperialistic, as well as requiring extra work for taxonomists. The development and implementation of ZooBank therefore need to be done in as user friendly a manner as possible, and registration needs to be made straightforward and simple. It will be rapidly obvious that the benefits of ZooBank will compensate for the effort and resources required for its development.

The establishment and general acceptance of ZooBank will depend partially upon the adherence of zoologists to the ICZN Code rather than adoption of any other proposed nomenclatural systems such as the Phylocode or Biocode. While the former appears not to have gained much acceptance since its introduction (De Queiroz and Gauthier, 1994), there is still a strong group of supporters for the Biocode, which in contrast does have a lot to recommend it. The reluctance of plant, fungal and animal taxonomists to radically change their *modus operandi* with respect to the code they follow suggests adoption of the Biocode is unlikely in the near future.

Web-based taxonomy will soon be a reality, and many of the aspects of ZooBank and the code that are affected by traditional publication will become irrelevant. A scenario whereby the act of registration would effectively constitute publication is clearly a strong possibility in the near future. Before that can happen, a rigorous and democratic peer-review system needs to be in place to enable solely web-based taxonomy, and some possible systems are discussed next.

POSSIBLE MODELS FOR ZOOBANK

Registration, publication and availability can be defined for our present purposes as follows: Registration is the process of entering a complete record in the ZooBank registry. Publication refers to code-compliant published works, as defined in Chapter 3 (Articles 7–9) of the fourth edition of the ICZN Code. An available name is a scientific name applied to an animal taxon that conforms to the provisions of the code.

Next we present the following three scenarios relating to registration and publication, and how they affect, and are affected by, the current code: 1: (Publication + Registration) = Availability; 2. Registration = Availability (Polaszek et al., 2005b); 3. Registration = Publication = Availability.

Scenario 1. Publication + Registration = Availability

To be available, names and acts must *both* be published in accordance with existing code rules, *and* be registered. Registration can take place either before publication or after publication. If before or within 2 years after publication, the date of availability is the publication date (Figures 8.1 and 8.2). If more than 2 years after publication, the date of availability is the registration date (Figure 8.3). The advantages of this scenario include relatively small changes to existing taxonomic practice, rapid implementation via an amendment to the fourth edition of the code, the maintenance of implicit quality control via traditional publication venues and, consequently, (perhaps) broader acceptance by the taxonomic community. Possible disadvantages include a somewhat complex procedure involving asynchronous publication and registration events, arbitrary time periods affecting dates of availability and petitions to the Commission in certain special circumstances.

However, given the existing complexities of the code, these procedures can hardly be considered as particularly complex. Another possible perceived disadvantage would be an ambiguous ‘grey zone’ between publication and registration when names and acts are ‘assumed’ to be available, even though technically not available until registered. Again, the probability is that most authors will register new names prior to publication, eliminating this problem entirely. While this scenario still suffers from all the complexities and ambiguities associated with traditional

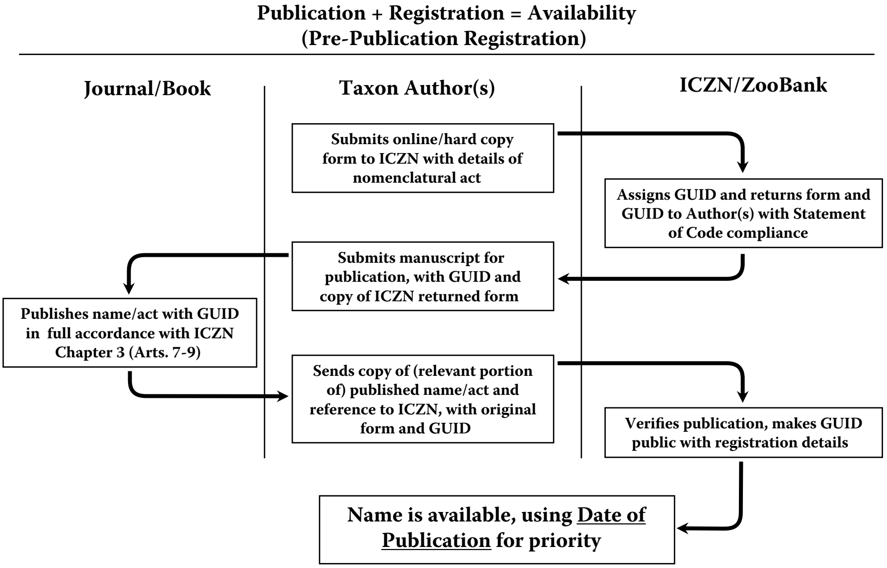

Publication + Registration = Availability
(Pre-Publication Registration)
Journal/Book
Taxon Author(s)
ICZN/ZooBank
Submits online/hard copy form to ICZN with details of nomenclatural act
Assigns GUID and returns form and GUID to Author(s) with Statement of Code compliance
Submits manuscript for publication, with GUID and copy of ICZN returned form
Publishes name/act with GUID in full accordance with ICZN Chapter 3 (Arts. 7-9)
Sends copy of (relevant portion of) published name/act and reference to ICZN, with original form and GUID
Verifies publication, makes GUID public with registration details
Name is available, using Date of Publication for priority

FIGURE 8.1 Publication + registration = availability (prepublication registration).

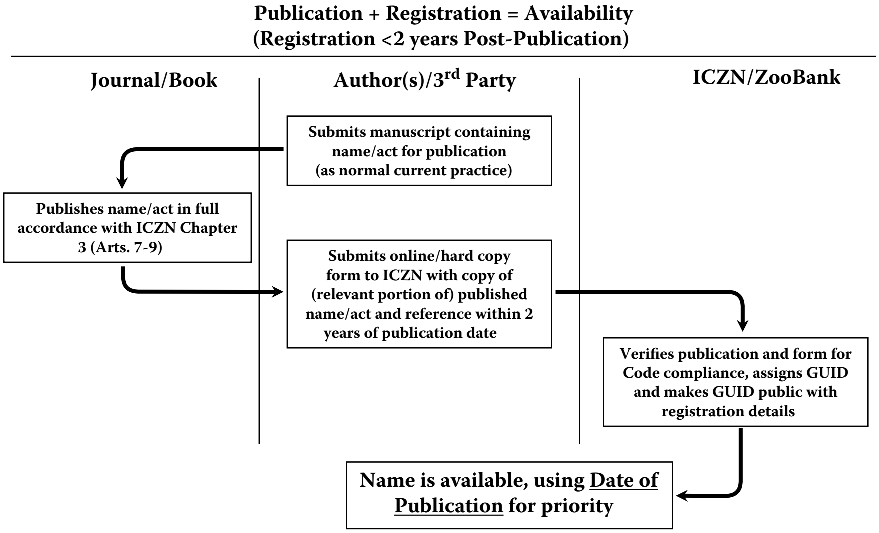

Publication + Registration = Availability
(Registration <2 years Post-Publication)
Journal/Book
Author(s)/3rd Party
ICZN/ZooBank
Submits manuscript containing name/act for publication (as normal current practice)
Publishes name/act in full accordance with ICZN Chapter 3 (Arts. 7-9)
Submits online/hard copy form to ICZN with copy of (relevant portion of) published name/act and reference within 2 years of publication date
Verifies publication and form for Code compliance, assigns GUID and makes GUID public with registration details
Name is available, using Date of Publication for priority

FIGURE 8.2 Publication + registration = availability (registration < 2 years postpublication).

paper publication entangled with nomenclatural availability, it would hardly differ from current practice, so would not really add up to an increase in complexity. Finally, scenario 1 would require an increase in the active role of ICZN staff (and associated costs) to process registration requests and verify code compliance before issuing LSIDs.

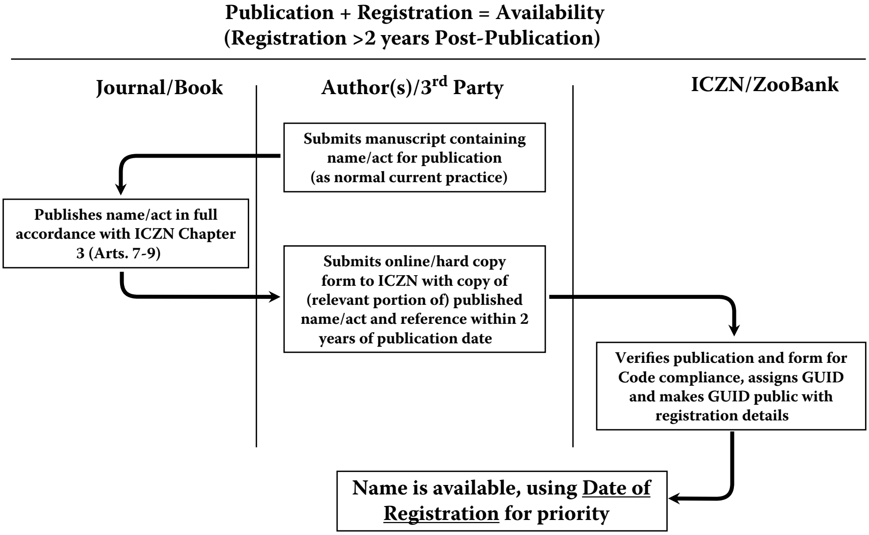

Publication + Registration = Availability
(Registration >2 years Post-Publication)
Journal/Book
Author(s)/3rd Party
ICZN/ZooBank
Submits manuscript containing name/act for publication (as normal current practice)
Publishes name/act in full accordance with ICZN Chapter 3 (Arts. 7-9)
Submits online/hard copy form to ICZN with copy of (relevant portion of) published name/act and reference within 2 years of publication date
Verifies publication and form for Code compliance, assigns GUID and makes GUID public with registration details
Name is available, using Date of Registration for priority

FIGURE 8.3 Publication + registration = availability (registration > 2 years postpublication).

Scenario 2. Registration = Availability

With this procedure, the process of registration itself is all that is required for availability of new names and acts. Prior or subsequent publication through traditional venues is encouraged, but is not integral to nomenclatural availability (Figure 8.4). Some advantages of this system would be that the legalities of nomenclatural availability and the science of taxonomy are disentangled from each other; there would be no ambiguity about dates of availability, and only minor increases to the active role of ICZN staff (and associated costs) would be required. Possible disadvantages include a fundamental change to the way taxonomic names and acts are established, eliminating the publication process from the act of nomenclatural availability. However, this would not necessarily be a problem from the perspective of the taxonomists (i.e. virtually the same as scenario 1), and in fact would only require a change to the technical legality of nomenclatural availability, not necessarily any change to taxonomic practice. To implement this system, more extensive changes are also needed in the code, such that these could probably only be implemented in a fifth edition. However, it will probably anyway take several years to work out the details and demonstrate the feasibility via a working voluntary registration system.

Another possible objection is that taxonomists would lose their primary benchmark for establishing professional status; that is, their CVs would have fewer publications listed. Taxonomists' professional status is established by publishing articles on scientific taxonomy and classification, which would continue exactly as before. Only the legalities of nomenclature would be dissociated from publications–not the science of taxonomy. While it is possible that some journals might not want to publish taxonomic descriptions if articles no longer carry the 'prestige' of establishing new names and acts in accordance with ICZN rules, it is also true that prestige in scientific publications comes from the quality of the science content of the published

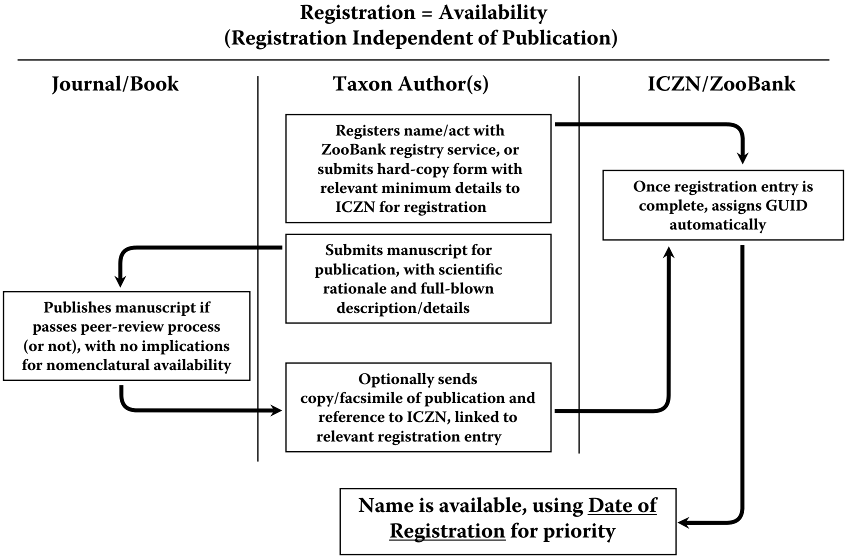

Registration = Availability
(Registration Independent of Publication)
Journal/Book
Taxon Author(s)
ICZN/ZooBank
Registers name/act with ZooBank registry service, or submits hard-copy form with relevant minimum details to ICZN for registration
Once registration entry is complete, assigns GUID automatically
Submits manuscript for publication, with scientific rationale and full-blown description/details
Publishes manuscript if passes peer-review process (or not), with no implications for nomenclatural availability
Optionally sends copy/facsimile of publication and reference to ICZN, linked to relevant registration entry
Name is available, using Date of Registration for priority

FIGURE 8.4 Registration = availability (registration independent of publication).

articles, and not from fulfilling a legalistic technicality for nomenclatural availability. Elimination of quality control/peer review from the process of establishing new names and nomenclatural acts could also be perceived as a disadvantage, but since the code requires neither peer review nor quality control, the scenario would be no different from the current situation. It could also be argued that the ICZN requirement for publication de facto forces most names and acts through peer review anyway. The possibility that bad taxonomists (and non-taxonomists) might abuse the system by registering hundreds of bogus and unneeded names, perhaps for unscrupulous reasons (e.g. selling names for money), is also unaffected by the choice of possible scenarios (i.e. it always remains possible). The same goes for those taxonomists who might never get around to publishing the full description after the name is registered, potentially creating many names without robust taxonomic definitions.

Scenario 3. Registration = Publication = Availability

Under scenario 3, ZooBank would host a comprehensive, edited and peer-reviewed online journal (such as *Zootaxa*) in which *all* names and acts must be published, similar to the situation with bacterial names. The science of taxonomy becomes part of the nomenclatural process (by changes to the code), and submitted manuscripts are open to non-anonymous review by any interested or concerned taxonomist (Figure 8.5); these individuals would be invited by the alerting service to review papers in their area of interest. Major advantages of this procedure include zootaxonomic publications appearing in a single venue instead of scattered across hundreds of journals, and the prevention of unscrupulous authors' 'stealing' by trying to submit plagiarized work to a journal that has a faster turnaround time. All manuscripts

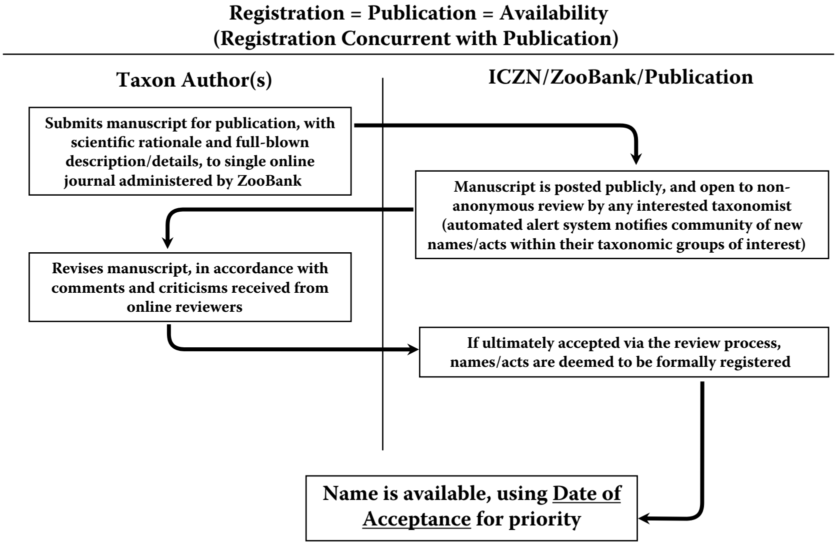

Registration = Publication = Availability
(Registration Concurrent with Publication)
Taxon Author(s)
ICZN/ZooBank/Publication
Submits manuscript for publication, with scientific rationale and full-blown description/details, to single online journal administered by ZooBank
Manuscript is posted publicly, and open to non-anonymous review by any interested taxonomist (automated alert system notifies community of new names/acts within their taxonomic groups of interest)
Revises manuscript, in accordance with comments and criticisms received from online reviewers
If ultimately accepted via the review process, names/acts are deemed to be formally registered
Name is available, using Date of Acceptance for priority

FIGURE 8.5 Registration = publication = availability (registration concurrent with publication).

would be examined by a large contingent of reviewers, instead of just a handful, thus greatly improving the reviews as well as democratizing the process. The reviews would also be public, instead of anonymous, so personal grudges or biases of the reviewers would be exposed to scrutiny by the whole community. Furthermore, a dedicated nomenclatural journal would mean that the review criteria will explicitly address all necessary aspects of code compliance and proper nomenclature. Other advantages of an online review process include speed and openness to feedback. Above all, copyright issues would cease to be a problem. This scenario would, of course, represent a major and fundamental change to the way taxonomy is done, both in terms of legalities of nomenclature as well as for the science of taxonomy.

With such major changes come certain complications, but the trade-off may well be worthwhile. With respect to online peer review, it must be borne in mind that many taxonomic groups do not have many (or even any) experts who might serve as reviewers, and thus submitted manuscripts may never receive peer review. This problem is equally true for traditional publication venues as well, but with only one 'official' taxonomic journal with potentially thousands of regular contributors and readers, there is a much better chance of finding someone who is qualified to review the manuscript. As with scenario 2, more extensive changes to the code would be required, such that it could probably only be implemented in a fifth edition of the code, perhaps several years in the future. In any case, it will probably take a long time to work out the details and demonstrate the feasibility via a working voluntary registration system. It could be argued that such a system would impose a huge burden on the taxonomic community to provide peer reviews to 20,000+ new names each year, but in fact the burden would be no more than that which already exists. For every manuscript submitted and reviewed through the official ZooBank online journal, one fewer manuscript would be submitted to a traditional journal, so there would be no net increase in the total number of manuscripts to review. A common argument against such a scenario is that existing journals that depend on taxonomic descriptions and nomenclatural acts to fill their pages and maintain a subscriber base may be driven out of business. However, it should be borne in mind that journals exist to serve scientists, not the other way around.

Criteria for determining when a submitted manuscript should be deemed 'accepted' and when (and by whom) will always be a subjective and contentious issue. This problem could be largely solved by having each manuscript assigned to an impartial 'referee' whose speciality is *outside* the particular taxon involved, and who is fully familiar with the code–serving the same role as a journal editor. Finally, the legalities of nomenclatural availability, and the subjective science of taxonomy, would, for the first time, be formally coupled under code rules. Controversial as this sounds, it may be that a significant proportion of zoologists feel that quality control and peer review *should* be part of the code's requirements for nomenclatural availability.

SUMMARY

Recent advances in science and technology have set the stage for an unprecedented opportunity to increase the speed and efficiency with which we are able to describe biodiversity. However, in order to maintain data quality, a monitoring system for the scientific names of animals far more effective than anything currently in place needs to be available. A mandatory register, proposed here as ZooBank, would provide such a system. Several scenarios for the running of ZooBank are possible, and these have been described in detail. The most radical, in which registration, publication and availability are all linked by having ZooBank effectively equivalent to a solely online journal, could provide the most democratic and efficient system, but would require extensive changes to taxonomists' current working practices, which may not be such a bad thing. All scenarios show many similarities with the GenBank system for archiving gene sequences on the web, and while GenBank is in theory a voluntary system, the insistence of most reputable journals for listing GenBank accession numbers creates what is effectively a mandatory register.

In this respect it may be desirable in the future to explore ways in which ZooBank and GenBank are more closely associated. This could mean more than simple web links between names and LSIDs, but rather a situation where GenBank and ZooBank are effectively the same. Concerns regarding solely digital archiving are also well illustrated by GenBank. While paper archiving has been shown repeatedly to be no guarantee of safeguarding data in perpetuity, at the same time it is inconceivable that the data that form the core of GenBank will ever be lost. With such valuable information representing the efforts of thousands of scientists, it is in the interests of everyone to safeguard that information. We believe that precisely the same argument holds for the millions of entries that will eventually form ZooBank–the complete register and database of all the scientific names of animals.

REFERENCES

CBD (2005) *Handbook of the Convention on Biological Diversity Including Its Cartagena Protocol on Biosafety,* 3rd ed., CBD/UNEP, pp. 1–1493.

De Queiroz, K. and Gauthier, J. (1994) Towards a phylogenetic system of biological nomenclature. *Trends in Ecology & Evolution,* 9: 27–31.

Dubois, A. and Nemésio, A. (2007) Does nomenclatural availability of nomina of new species or subspecies require the deposition of vouchers in collections? *Zootaxa,* 1409: 1–22.

Godfray, H.C.J. (2002a) How might more systematics be funded? *Antenna,* 26(1): 11–17.

Godfray, H.C.J. (2002b) Challenges for taxonomy. *Nature,* 417: 17–19.

ICZN (1999) *International Code of Zoological Nomenclature*, 4th ed., International Trust for Zoological Nomenclature, London, pp. 1–306.

Polaszek, A., Grubb, P., Groves, C., Ehardt, C.L. and Butynski, T.M. (2005a) Response to Landry and Timm et al. *Science,* 308: 1161–1164.

Polaszek, A., Agosti, D., Alonso-Zarazaga, M., Beccaloni, G., Bjørn, P., Bouchet, P., Brothers, D.J., Cranbrook, G., Evenhuis, N., Godfray, H.C.J., Johnson, N.F., Krell, F.-T., Lipscomb, D., Lyal, C.H.C., Mace, G.M., Mawatari, S., Miller, S.E., Minelli, A., Morris, S., Ng, P.K.L., Patterson, D.J., Pyle, R.L., Robinson, N., Rogo, L., Taverne, J., Thompson, F.C., van Tol, J., Wheeler, Q.D. and Wilson, E.O. (2005b) A universal register for animal names. *Nature,* 437: 477.

Polaszek, A., Alonso-Zarazaga, M., Bouchet, P., Brothers, D.J., Evenhuis, N., Krell, F.-T., Lyal, C.H.C., Minelli, A., Pyle, R.L., Robinson, N., Thompson, F.C. and van Tol, J. (2005c) ZooBank: The open-access register for zoological taxonomy, technical discussion paper. *Bulletin of Zoological Nomenclature,* 62(4): 210–220.

Thorne, J. (2003) Zoological record and registration of new names in zoology. *Bulletin of Zoological Nomenclature,* 60(1): 7–11.

Wakeham-Dawson, A., Morris, S., Tubbs, P., Dalebout, M.L. and Baker, C.S. (2002) Type specimens: Dead or alive? *Bulletin of Zoological Nomenclature,* 59(4): 282–286.

Wilson, E.O. (2003) The encyclopedia of life. *Trends in Ecology & Evolution,* 18: 77–80.

9 Understanding Morphology in Systematic Contexts

Three-Dimensional Specimen Ordination and Recognition

Norman MacLeod

CONTENTS

INTRODUCTION

There are many definitions of morphometrics, but perhaps the broadest is 'the associations, causes, and effects of form' (Bookstein, 1991). Over the last 50 years morphometric methods have been instrumental in advancing the study of how organismal shape covaries with size (allometry), time (evolution), environment (ecophenotypy), geography (morphological biogeography), ecology (ecomorphology), development (heterochrony), and a host of related factors (see Sokal and Sneath, 1963; Blackith and Reyment, 1971; Reyment, 1971; Sneath and Sokal, 1973; Pimentel, 1979; Reyment, 1980; Reyment et al., 1984; Reyment, 1991). More recently, the school of geometric morphometrics[1] has added further rigor to this research programme by synthesizing deformational (e.g. Siegel and Benson, 1982) and multivariate (e.g. Blackith and Reyment, 1971; Bookstein, 1978) approaches to shape comparison and placing both on a firmer mathematical footing (see Kendall, 1977, 1984; Mardia and Dryden, 1989; Rohlf and Bookstein, 1990; Bookstein, 1991). Books, technical articles, conference abstracts and student theses that employ morphometric approaches now abound (Adams et al., 2004). How odd then that morphometrics in general and geometric morphometrics in particular have largely neglected the most fundamental biological correlate of form–the correlation on which virtually all other morphometric studies critically depend and that provides the most obvious link between morphometrics and mainstream biology: the correlation between form and taxonomy.

As anyone who works in the field of systematics already understands, the overwhelming majority of species, genera, families and other categories are practically recognized through reference to some aspect of their morphology. Many species concepts have been advanced over the years (see Wheeler and Meier, 2000, for a recent review). These can be subdivided into pattern-oriented concepts (e.g. typological species, morphological species) that focus on how species are delineated, and process-oriented concepts (e.g. biological species, ecological species, phylogenetic species) that focus on the ways cohesion within, and distinctions between, species originate and are maintained (see de Queiroz and Donoghue, 1988, 1990; Wheeler and Nixon, 1990, for additional discussion). But no matter what criteria are used to define or explain the existence of species, their day-to-day identification is overwhelmingly undertaken as an exercise in comparative morphology. This chapter will consider whether taxonomy and systematics can benefit in any substantive way from the application of morphometric analysis at present and what challenges need to be overcome to make the morphometric toolkit more useful to taxonomists–systematists of the future.

Of course there are those who, impressed by the potential of DNA sequences as a means of recognizing species' 'barcodes' (e.g. Hebert et al., 2003; Tautz et al., 2003; Hebert et al., 2005), suggest we are on the cusp of a revolution in the manner in which all species will be recognized. Despite a number of optimistic studies (e.g. Zhi et al., 1996; Chu et al., 1999, 2003; Sun et al., 2003; Tang et al., 2003;

Shih et al., 2004), and the systematics of a few morphologically conservative groups aside (e.g. fungi), both the primacy and practicality of morphological observation as the most reliable generalized identification method is, for the vast majority of organismal groups, unlikely to change in the foreseeable future. Molecular approaches may work for some groups, but the current consensus among systematists is that molecular methods will not be able to provide the information necessary to identify the majority of modern species reliably (see Wheeler, 2003; Wheeler et al., 2004; Ebach and Holdrege, 2005; Cameron et al., 2006; Hickerson et al., 2006; Meier et al., 2006; Wheeler, 2007, this volume and references therein), much less ancient species whose identification is just as important to systematics (see Donoghue et al., 1989; Smith, 1994). Therefore, the contribution morphometric approaches can make to the question of group identification is highly pertinent to the success of the entire systematic enterprise, both now and in the future.[2]

Development of new methods to support accurate taxon identifications is also of the utmost importance. It is by now an open secret that the world is running out of qualified taxonomists who are competent to identify the very biodiversity the world is losing at an ever more alarming rate (Kaesler, 1993; Wheeler, 2007, this volume). There are many reasons for the reduction in qualified taxonomists, but all come down to a simple matter of perceived costs and benefits. Taxonomy is perceived to cost a lot to do right.

In addition to the salaries of taxonomists and their support staff, taxonomic research has traditionally required large specimen collections and large libraries to maintain appropriate reference material for morphological investigations. Since the 1950s taxonomists have also increasingly required access to advanced sensing and analytical facilities (e.g. scanning electron microscopes, DNA sequencing laboratories, CT scanners). Taxonomy is not as expensive as other branches of science, but the expenses involved are not inconsiderable, especially if research centres need to make large numbers of identifications for a wide range of organismal groups. Indeed, the only research centres that can comprehensively fulfill this remit at present are the large research-oriented, natural history museums, of which there are only a few.

Few argue that taxonomy is irrelevant to modern science (see the 2003 UK House of Lords Report on systematics, *What on Earth*). But there is a common perception that taxonomists–systematists have not taken sufficient advantage of technology to improve their productivity and standardize the quality of their product (see Godfray, 2002). In many industrial settings (e.g. agriculture, medicine, pharmaceuticals, petroleum exploration) taxonomy is often not regarded as a 'front line' or 'first recourse' approach to problem solving because of the time it takes to collect specimens, transport them to the specialist (or the specialist to them) and for the specialist to make their identifications that are usually communicated back to the corporate centre in a form that can range from a simple list of taxa to a heavily qualified and referenced technical report. In addition, it is often the case that different commercial laboratories produce taxonomic identifications that disagree with one another, even when identifying relatively common and abundant taxa. Academic consultancies often produce marginally better results, but are not immune to these problems, usually are not equipped to provide large-scale analyses and are dwindling in number because progressively fewer university departments are maintaining organismal biology programmes (see Wheeler et al., 2004; Wheeler, 2005, for a discussion of the reasons behind this trend). Put badly, molecular research strategies are perceived to be able to do more with less investment of resources than morphology-based approaches.

The effect of this long-term decline in morphology-based taxonomy–termed the 'taxonomic impediment'–has been that, just as society is coming to understand the full implications of global climate change, overpopulation, and widespread habitat destruction, taxonomic experts in all but the most charismatic groups are in danger of becoming as rare as some of the species they study. This situation has a detrimental effect on all biology and diminishes humanity's overall ability to understand, conserve and utilize the Earth's natural resources in a sustainable manner. What to do?

Ever since the industrial revolution, whenever humans have been confronted with the need to perform repetitive, well-defined tasks quickly, with a high degree of consistency and efficient resource usage, the preferred solution has been to automate. Automated species identification systems have been the stuff of science fiction for generations (see Janzen, 2004). In order to be practical such systems should be based on the identification of patterns of morphological variation since this is the way most species have been, are now and will continue to be identified. Accordingly, it is to morphometrics that systematics must turn to make the type of generalized progress in this area that is needed. Because of the practical problems resulting from the taxonomic impediment, the nature of the raw material that must be identified and the training/background necessary to marry emerging automation technology with the necessary data analysis approaches, it is difficult to imagine a group of researchers with more potential to contribute to the rejuvenation of taxonomy and systematics than morphometricians.

This chapter has been written in an explicit attempt to foster a renewed interest in the practical aspects of morphometrics and the automated species-group identification issue by all those who care about systematics and organismal biology. It seeks to provide a context within which working taxonomists–systematists might better understand recent developments in morphometrics and related fields of quantitative data analysis that may contribute to their biological goal of developing a better understanding of the number, origin and roles of organismal groups in nature. More than 25 years of conversations with colleagues and students across the biological disciplinary spectrum have convinced me there is a good deal of interest in these issues, but a dearth of understanding of the current state of the art, much less the most promising avenues for future research. To facilitate this understanding the chapter will be presented in two parts: an initial review of concepts from the general fields of morphometrics of pattern recognition followed by a presentation of a set of example analyses that bear on the questions of (1) whether biological groups can be recognized using morphometric methods and (2) which methods are best suited to which identification tasks.

Discussions of some multivariate approaches along these lines can be found in MacLeod (2007a,b). Those presentations focused on the collection and analysis of information contained in traditional two-dimensional morphometric data and images. By way of contrast, this chapter will focus on the collection and analysis of information contained in all three Euclidean dimensions. Whereas it is now trivially easy and inexpensive to collect two-dimensional images, three-dimensional scanning will be the preferred mode of representing biological specimens in the not-too-distant future.

Morphometrics is already making the transition from two to three dimensions now that accurate three-dimensional scanners and so-called three-dimensional printers are becoming available. The extension of morphometric data analyses approaches into the third dimension will also resolve ambiguities that are unavoidable in two-dimensional analyses and that qualitative morphological analyses take advantage of on a routine basis. While it should be emphasized that many interesting problems in morphological systematics can be addressed successfully with two-dimensional approaches, even more can be addressed using three-dimensional data and addressed in a manner that brings all the advantages of quantitative approaches to bear on a greater proportion of the biological data at hand. This chapter will also present a new morphometric data analysis technique–'eigensurface analysis' (see also MacLeod and Polly, 2006; Polly, 2006; Polly and MacLeod, in press)–that has proven useful in the contexts of both three-dimensional morphometric data analysis and automated group identification.

IMAGES, THREE-DIMENSIONAL SCANS, MORPHOMETRIC ORDINATION AND NEURAL NETS

Digital Images

For all specimen-identification applications discussed in this chapter, digital images and/or scans will serve as the source for quantitative data. Although it seems obvious and natural to regard images of specimens as an adequate substitute for actual specimens–after all systematists have used images to illustrate the characteristics of, and differences between, systematic groups for literally hundreds of years (see Rudwick, 1972)–a brief consideration of the advantages and disadvantages of digital images as a source of geometric information is instructive.

Digital images are essentially rectilinear grids that create an image out of cells (picture elements, or pixels) that are assigned a greyscale or color (RGB, CMYK) value (Figure 9.1). In a digital photographic image these cells are filled by values that match or mimic those of the original scene. The *x* and *y* axes of the grid provide an internal means of both characterizing overall form of objects within the scene or frame and of spatially locating these objects within the frame and relative to one another. This is not the case with analogue images, which contain spatial information but lack the internal, quantitative spatial-reference system set up by the matrix of pixels. The quantitative data necessary to allow both specimen and/or scene analysis are built into the very structure of digital images.

The greyscale or color values assigned to each cell or pixel of the image matrix represent not only the shade or color present at that location in the scene, but also aspects of the specimen's texture, composition and position relative to the ambient lighting source. With respect to assessing the specimen's three-dimensional shape, the latter aspect of this shade-color value is the most important. Unfortunately, in a normal image this information is complexly confounded with texture and

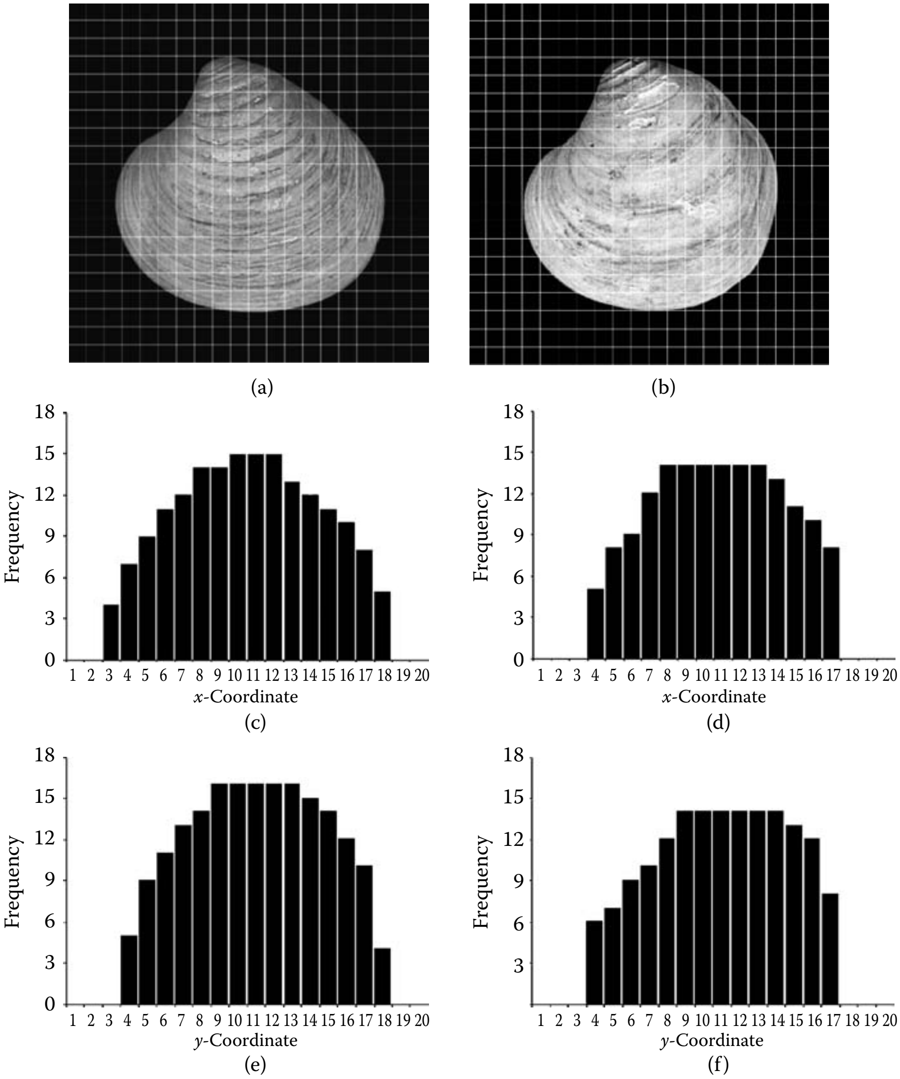


FIGURE 9.1 Images of two specimens of the bivalve *Astarte mutabilis* with a superimposed sampling grid representing the digitization process. In creating a digital image the scene is broken into a matrix of quantified brightness/color values: the picture elements or pixels. In addition to their image-representation properties, the pixel grid imposes a spatial coordinate system that can be used to quickly and easily characterize aspects of the entire scene–in this case the outlines of the specimens–as well as the relative locations of features or substructures. Here, the distributions of the number of pixels falling on the image are show for the columns (upper distributions) and rows (lower distributions) of the pixel matrix. These distributions could be used to quantify absolute dimensions of the specimens (e.g. relative fatness–thinness) as well as aspects of the outline shape. Morphometric image analysis systems, three-dimensional digital scanners and image-based ANNs are all based on the ability of digital images to represent shape within a convenient, internal spatial reference system.

compositional variations. This gives two-dimensional images (both analogue and digital) a spatial dimensionality greater than two, but less than three.

There is another aspect of digital images that is important to appreciate from the standpoint of the group ordination-identification problem that relates their structure to both geometric morphometrics and numerical systematics. The grid imposed on the scene during digital image capture can be regarded as a sampling grid composed of topologically corresponding units within the frame. So long as this grid system is tied to the image frame and not the specimen it is both conceptually and practically distinct from the landmark systems employed by geometric morphometrics. Nevertheless, intriguing similarities exist between this frame-based sampling system and phenetic resemblance concept embodied by ideas of 'numerical homology' (see Sokal, 1966; Sokal and Rohlf, 1966) and employed on a routine–though largely subconscious–basis by taxonomists undertaking qualitative assessments of morphology.

When geometers look at shapes they note differences in the distribution of points within the shape's boundary relative to an externally imposed coordinate system. For example, 'fat' shapes may have a greater proportion of intermediate coordinates included within the shape boundary (Figure 9.1) than 'thin' shapes, which have a smaller number of coordinates included along one (not both) of these axes. The conceptual distinction between 'fatness' and 'thinness' as represented by these characterizations is clear and its operational implementation in an automated shape recognition system is trivial. Under these definitions, fatness and thinness can be recognized irrespective of any biological property of the organism whose shape is being assessed.[3] It is simply a matter of geometry.

To systematists reading this chapter the foregoing description should sound vaguely familiar. Systematists have employed this approach to shape assessment for literally thousands of years. To take a familiar–though slightly oblique–example, consider the assessment of character states. A systematic character is a putatively homologous aspect of the organism's phenotype shared among a group of organisms (see Colless, 1985; Firstrup, 1992; Scotland and Pennington, 2000; Humphries, 2002, and references therein). Testing character homology is accomplished via phylogenetic analysis and so requires an assessment of cladistic (as opposed to phenetic) similarity (see Kitching, 1998). Character states are different, though. Ontologically, character states exist below the level of homology. They are simply different variants of the form of a set of structures that are all at least putatively homologous by definition.[4] Accordingly, character states are recognized on the basis of phenetic similarity in the manner described earlier. A feature of an organism may be described as fat or thin because of its relative proportions (Figure 9.1). Similarly, it can be described as long or short, compressed or expanded, simple or complex, etc. These sets of phenetically delineated character states have been employed routinely throughout systematics. But what do character-state distinctions like fat or thin, long or short, etc. really mean? Sometimes rules of thumb are provided (e.g. >50%, <50%), but these are rarely tied to any quantitative data describing or justifying them.

Since character states are operationally defined on the basis of phenetic distinctions (e.g. fat–thin, tall–short, simple–complex), morphometric approaches should be used to assess the nature of the putative character's distribution (continuous or clustered into discrete modes or ‘phena’) and confirm the appropriateness of the character state assignments (MacLeod, 2002a). That such morphometric testing rarely happens in practice is an outstanding anomaly within contemporary phylogenetic systematics that compromises its much referred to objectivity (see preceding references) and degrades its acuity insofar as incorrect or arbitrary character-state assignments can lead to erroneous phylogenetic results. For the purposes of the current argument, however, the important point is that systematists inevitably employ a phenetic approach to the assessment of character states and that this assessment is made in accordance with a concept of recognizing topological correspondences by qualitatively scanning across the structure in question and obtaining a ‘sense’ or estimate of its shape. Not only are character states identified/recognized in this phenetic manner, so are the combinations of character states that systematics calls ‘species’.

There are no rules for the recognition of species. Indeed, there are not even any tests that can be applied universally. Smith (1994) describes the species-recognition process as beginning with the recognition of overall shape-based phena: collections of organisms that appear to exhibit a regularity of (usually morphological) type that allows a group to be recognized objectively and distinguished from other such groups. These phena are confirmed by identifying the characteristics that give the group its distinction and formally elevated to species status simply by describing these characteristics in a particular type of publication series. Such descriptions provide textual summaries of the overall nature, and limits, of variation in the attributes characterizing these phena. Tests of species concepts can be made either at the gross morphological level (e.g. morphological species concept) or the character-state level (phylogenetic species concept), with the latter being incorporated into subsequent phylogenetic analysis. As before, the point is that, in practice, these represent geometric comparisons conducted in a phenetic mode. Since species are identified on the basis of character-state criteria and character states are identified via phenetic comparison, species must be regarded as phenetically recognized entities.

Once species have been designated, phylogenetic analysis can assess the structure of historical relations among species and test statements of homology (including the identification of apomorphies). At this stage particular species concepts may be associated with phylogenetically unique character states. But this recognition happens after a phenetically based species concept has been created, not before. Also, in an evolving lineage it will often be the case–at least in principle–that objectively recognizable and practically useful segments of evolving lineages will not posses any truly unique character states (e.g. ancestors; see Eldredge and Cracraft, 1980). These lineage segments *must* be ontologically defined and practically recognized using phenetic criteria. In sum, a commitment to phylogenetic analysis *per se* can neither correct for inappropriate species designation nor displace phenetics from its necessary role in both character-state and species identification.

If assessing phenetic similarity is an irreducible part of systematic practice for either character or species recognition, systematists–taxonomists must consider how best to take advantage of this fact. As demonstrated previously, digital assessments of specimen morphology can sample and summarize all aspects of the topological information needed to support phenetic similarity tests. But in terms of generalized size and shape analyses, the use of two-dimensional images is problematic because of the absence of high-quality information about morphological variation in the direction of the optic axis (= the morphometric *z*-direction).

Prior to the advent of digital image analysis systems (see MacLeod, 1990; Becerra et al., 1993), attempts were made to incorporate some aspects of three-dimensional data collection in many morphometric studies (e.g. via interlandmark distance measurements made with callipers). These attempts were largely confined to specimens that could be manipulated easily by hand. Certain biological structures also lend themselves to two-dimensional characterization (e.g. bee wings, butterfly wings, leaves). Although digital images do portray aspects of the third dimension, and regardless of the fact that this information is used to place landmark points and trace outline curves, the data that result from these placements and tracings remain strictly two dimensional in the majority of morphometric studies. Accordingly, it must be suspected that a strictly two-dimensional morphometric data collection approach based on digital (or analogue) images will not capture quantitative data comparable to the amount and the nature of the data used by qualitative taxonomists to recognize the vast majority of species. One solution to this dilemma may be to abandon reliance on digital imaging to represent species morphologies and make greater use of three-dimensional surface scanning technology.

Three-Dimensional Scans and Virtual Specimens

Although three-dimensional scanners have been available for some time, these devices are only now making their way out of engineering and design laboratories and into both museums and university biology departments. A wide variety of models are currently available that employ a variety of technological approaches to the collection of three-dimensional coordinates (e.g. laser scanning, stereophotometry, interference grids, magnetic positioning, acoustic positioning). Costs vary from a few thousand to over a million GBP. Spatial resolutions of 0.01 mm are able to be achieved even by low-middle-priced systems.

Three-dimensional scanning systems usually output scaled *x,y,z* coordinate values–or vertices–in the form of a point cloud (Figure 9.2a). These coordinates can be exported to ASCII files that can serve as input to various mathematical data-analysis routines (e.g. Procrustes superposition, PCA, SVD). Alternatively, three-dimensional coordinates can be used to create a representation of the specimen on the computer's monitor. This is accomplished by joining the coordinates with chords to simulate a polygonized surface (Figure 9.2b) and then filling the polygon areas with a contrasting color (Figure 9.2c). Once polygonized, the virtual surface can be standardized for orientation and lighting quickly and easily (Figure 9.2d).

It is often the case that a single three-dimensional scan of a specimen is not sufficient to represent all aspects of the morphology. To overcome this problem most three-dimensional data manipulation software include either manual or automated procedures for merging or registering different scans of the same object into a composite. In the manual mode this operation is accomplished by locating at least three common points on the different scans and then superimposing them, often using a variant of the Procrustes superimposition algorithm. Once the equations of the landmark-point superposition have been determined these can be used to merge an entire

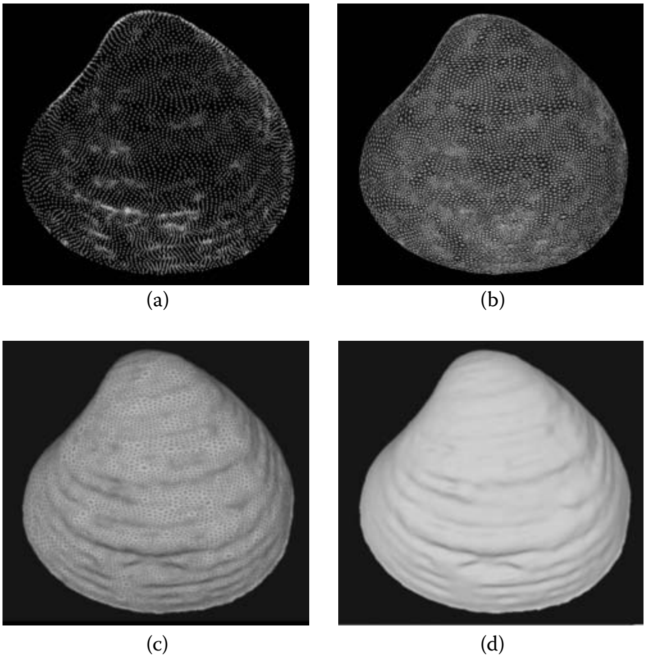


FIGURE 9.2 Steps in creating a virtual object from a three-dimensional surface scan. (a) Cloud of *x,y,z* points collected by the three-dimensional scanner. (b) Wireframe mesh of points interconnected into a surface composed of tessellated polygons. (c) Filled virtual surface formed by assigning color and brightness values to each polygon. Note interpolygon brightness/color values have been adjusted to simulate a light source in the upper left corner of the frame. (d) Finished virtual object whose surface has been smoothed and both the original points and polygon boundaries removed. High-quality virtual images like this can be constructed quickly and easily given appropriate–and increasingly inexpensive–hardware/software. Such data and images have many uses throughout systematics.

point cloud onto the target point cloud. When automated, this procedure involves marking of the reference and target points prior to scanning, tracking the spatial relations between scans (e.g. angles between scans for specimens mounted on a turntable) or recursively searching various superimposition patterns until the best fit is achieved. It is usually necessary to employ a combination of automated and manual registration to obtain a complete virtual object. Once an appropriate composite scan has been completed or virtual object constructed, the coordinate data and/or images of the specimen in a standardized pose and with standardized lighting effects can be obtained, exported (if necessary) and used in subsequent analyses.

Morphometric Ordination

There are several schools of thought regarding morphometric technique and at least two schools of thought regarding the best way to approach the automated group-identification problem. Accordingly, a brief history of these schools is needed to understand underlying conceptual similarities and differences, as well as to clarify general points about future technique-centred research directions.

Morphometrics has been used to characterize groups of organisms at least since Aristotle made the observation that 'there are certain things which suffer no alteration (save of magnitude) when they grow' (see Thompson, 1917). D'Arcy Thompson was among the first to call attention to the distinction between size and shape in organisms and portions thereof, drawing on sources as diverse as Kant, Newton, Da Vinci and Dürer for examples. Thereafter, Huxley (1932) developed mathematical formulae for expressing this distinction, including what came to be termed the allometric equation (Huxley and Teissier, 1936; see Gayon, 2000, for a review). While originally an exercise in bivariate data analysis, the concept of allometry was extended to multivariate data-sets by Jolicouer and Mosimann (1960) (see also Jolicouer, 1963). Of more direct application to the group-identification problem, R.A. Fisher (1936) developed a canonical method for optimizing the difference between two sets of measurements, illustrating this method using a biological group-discrimination problem in the form of the (now) famous *Iris* data-set (see MacLeod, 2007a,b, for an extended discussion). This was followed by the formulation of growth-free discriminant functions by Burnaby (1966) and Gower (1976).

By 1971 Blackith and Reyment had christened the emerging field of mathematical size–shape analysis 'morphometrics', which was being referred to as the next phase of the (then) popular numerical taxonomy movement (see Sokal and Sneath, 1963; Sneath and Sokal, 1973). Numerical taxonomists were intrigued by the idea of automated scanning for the purposes of identification and classification. Sokal and Rohlf (1966) performed experiments with group identification based on the random sampling of morphology and Sneath (1967) sought to quantify D'Arcy Thompson's 'transformation grid' approach to shape analysis through application of trend surface analysis (Figure 9.3). Numerical taxonomists clearly anticipated later developments in morphometric theory when they pointed out that, whereas non-linear patterns of large-scale spatial variation might be responsible for generalized (= phenetic) resemblance, it was often the case that patterns of small spatial-scale (non-linear) variation carried the more important taxonomic signal, at least in terms of species identification (see Sneath and Sokal, 1973). Unfortunately, this second (morphometric) wave of the numerical taxonomic movement was set back when a schism developed between those taxonomists who preferred phenetic and those who preferred cladistic approaches to interobject similarity assessment (see Hull, 1988).

Almost a decade later research groups led by Bookstein (1978, 1980), Reyment (1980), Strauss and Bookstein (1982), Siegel and Benson (1982), Ehrlich et al. (1983), Lohmann (1983), Reyment et al. (1984) and Benson et al. (1982) were all pursuing what appeared to be quite different numerical approaches to the quantification of organismal shape variation. However, in the late 1980s these approaches were suddenly, and somewhat unexpectedly, synthesized–largely through the efforts of Fred Bookstein (1986, 1990, 1991), but with important contributions by David Kendall (1984), Kanti Mardia and Ian Dryden (1989a,b), Colin Goodall (1991), and F. James Rohlf (Rohlf and Slice, 1990; Rohlf, 1993). This work resulted in formulation and subsequent development of geometric morphometrics.

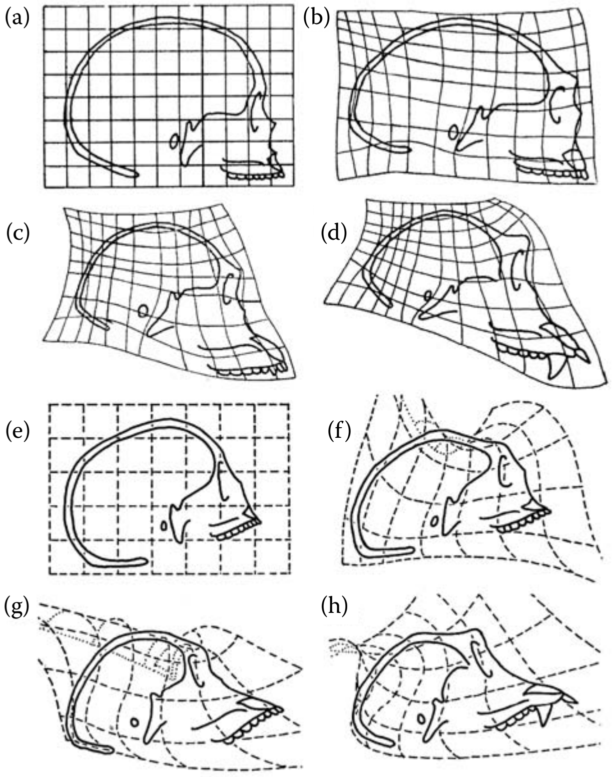


FIGURE 9.3 Trend surface analysis of four primate skulls illustrating the first attempt to quantify the transformation grid concept. (a) E. *Homo*; (b) F. *Pithecanthropus*; (c) G. *Australopithecus*; (d) H. *Pan*. A-D hand-drawn transformation grids. (e–h) trend-surface interpolations. Adapted from Sneath (1967).

Geometric morphometrics is constructed on the concept of shape coordinates. Given a set of shapes defined by a series of Euclidean *x,y* coordinates placed at corresponding 'landmarks' across a sample of biological forms[5] (Figure 9.4a), the landmark-defined shape of each form can be compared by superimposing the centroids of each landmark set and then scaling and rigidly rotating the landmark configurations of each relative to a reference shape (usually the mean shape of the sample) until the sum of squared differences between corresponding landmarks is minimized (Rohlf and Slice, 1990; Bookstein, 1991; see Figure 9.4c).[6] Because this operation involves translation, scaling and rotation operations, it has been likened to a Procrustean process and is termed Procrustes superimposition.

Following Procrustes superimposition similarities and differences between sets of landmark vertices can be expressed in one of two ways: as the pair-wise covariance matrix of corresponding shape coordinates or as the sum of squared residuals from a reference configuration (usually the mean shape; this metric is the partial Procrustes distance). From either of these data a shape covariance matrix can be used to quantify geometric similarities–differences and principal components analysis

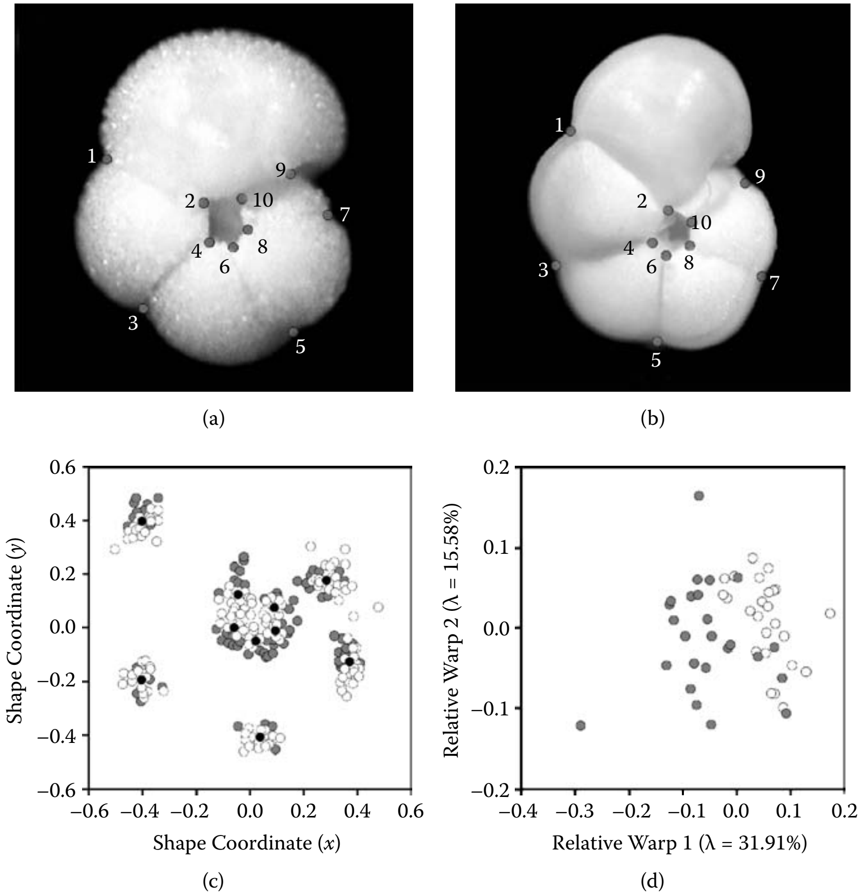


FIGURE 9.4 Example of a relative warp analysis of two planktonic foraminifer species. (a) *Globigerinella siphonifera* with positions of 10 landmarks shown. These represent Type 1 landmarks (see text) that quantify aspects of the chamber sequence, the shell coil, and the umbilicus size position. (b) Topologically corresponding landmarks shown on a specimen of *Globorotalia tumida*. (c) Shape coordinates for the 10 clouds of landmark positions after Procrustes superposition (GLS) of a combined sample of 27 *Ge. siphonifera* (white symbols) and 27 *Gr. tumida* (grey symbols) specimens. The black symbols represent the mean shape for the pooled sample. Note the former species appears to be characterized by a slightly tighter coil and larger umbilicus. (d) Scatterplot of the original specimens in the space formed by the first two relative warps. Together these axes represent 47 per cent of the observed shape variation. Note the broad differences in the distribution of *Ge. siphonifera* (white symbols) and *Gr. tumida* (grey symbols), indicating consistent differences in those aspects of the shape measured by the landmarks. Note also that, despite clear and numerous differences in the morphologies of these two species, the relative warps analysis results suggest considerable morphological intergradation.

(PCA) or singular value decomposition (SVD) employed to reorder the shape-coordinate covariance matrix into its major and minor modes of linear variation. The result–termed relative warps analysis–is usually expressed as an ordination of shapes

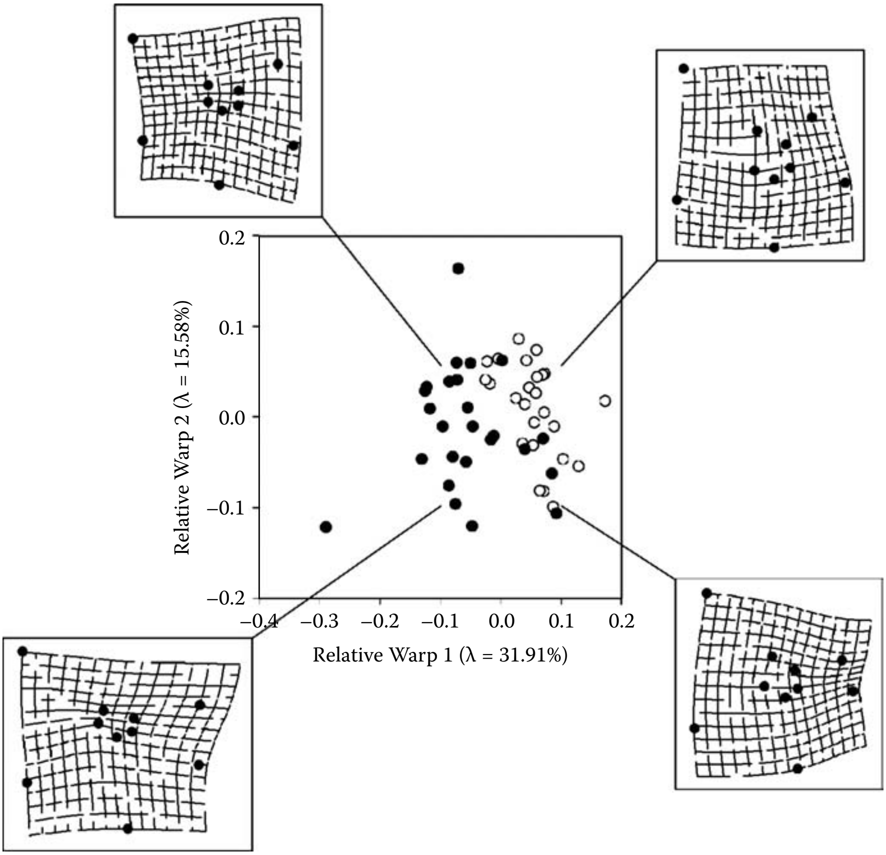


FIGURE 9.5 Relative warps scatterplot shown in Figure 9.4d with superimposed thin-plate spline models calculated for the coordinate positions indicated (black symbols). Inspection of these theoretical shape models confirms the speculation (based on Figure 9.4c) that *Gr. tumida* (grey symbols) is characterized by a smaller umbilicus and a looser coil than *Ge. siphonifera* (white symbols). See Figure 9.4 caption for symbol conventions.

within the linear space of the first few component/singular axes (Figure 9.4c). If the latter formulation is preferred, the implicit deformations can also be summarized as a displacement of each landmark along the z-axis (for two-dimensional data) with the deformation's overall geometry being expressed as though it were a steel plate bent or warped in the regions of relatively higher displacement from the reference configuration. This quasi-graphical convention is called a thin plate spline (Figure 9.5).

Using the thin plate spline as a physical metaphor for shape change, the geometry of the entire deformation across all landmarks can be represented by a 'bending energy matrix' (Bookstein, 1989, 1990, 1991; Rohlf, 1993). This index expresses the amount of curvature of the warped surface at each landmark location in the x and y (and z in the case of three-dimensional data) dimensions relative to a reference configuration. Once the bending energy matrix of a deformation has been determined its geometry can be decomposed into set of linear (= isotropic) and non-linear

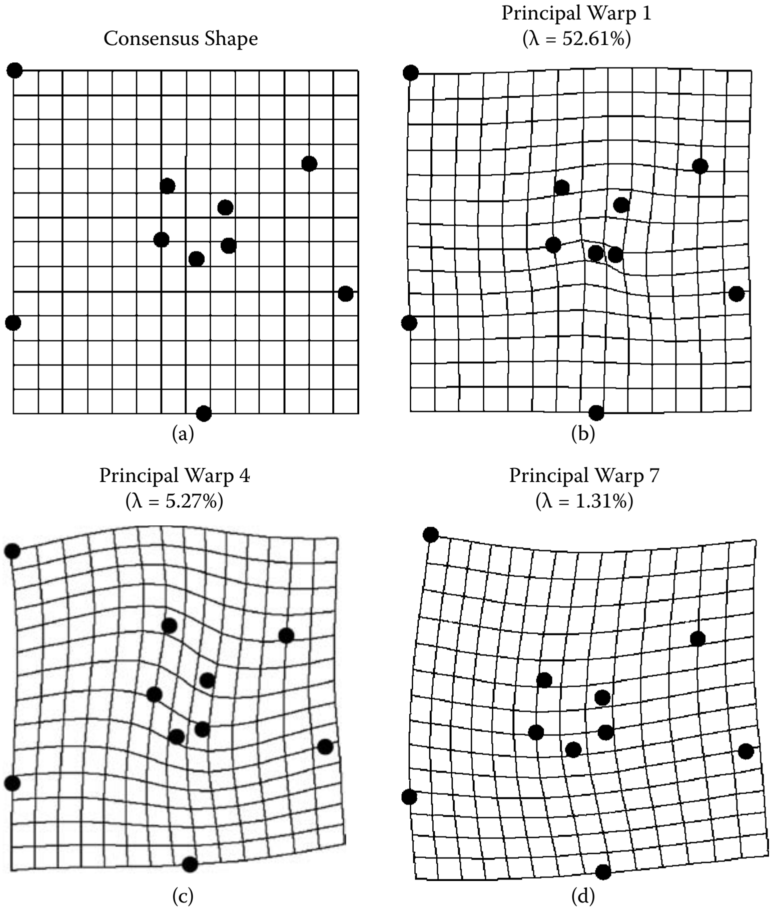


FIGURE 9.6 Principal warps for the *Ge siphonifera–Gr. tumida* sample. Consensus shape (upper left) represents mean shape for the combined sample. Principal warp 1 (upper right) represents the non-linear deformation that accounts for the greatest bending energy across the sample. Note this is a relatively localized deformation spatially centred on the umbilical region. Note also that landmarks located on the test periphery are also involved in this deformation mode, albeit to a lesser extent. Principal warp 4 (lower left) represents an intermediate non-linear deformation involving changes to both the umbilical area and the text periphery (= coil tightness and symmetry). Principal warp 7 (lower right) represents the non-linear deformation that accounts for the least bending energy. Note this deformation represents a translation of the relatively undeformed umbilical region laterally towards the penultimate chamber. This shape change is consistent with an asymmetric narrowing of the overall test.

(anisotropic) patterns (Figure 9.6) with the latter further decomposed into modes of shape variation that vary from the spatially localized to the spatially generalized. These are termed principal warps (Bookstein, 1991), and the x and y (and z in the case of three-dimensional data) components are termed partial warps.[7]

Unlike the PCA or SVD analyses of the covariance matrix between Procrustes-aligned shape coordinates, the sets of principal/partial warps for a sample of objects are best understood as a way of re-expressing the manner in which each shape differs from the reference configuration. In other words, the principal/partial warps contain the same geometric information as the original shape coordinates, albeit information that has been reorganized into a more geometrically structured format. This equivalence is demonstrated by subjecting the entire set of principal warp scores to a PCA/SVD analysis, in which case the results will be identical to a relative warps analysis (see earlier discussion). Both principal warps analysis and relative warps analysis can make use of the thin plate spline (Bookstein, 1989, 1991) to portray their results in a manner reminiscent of Thompson's transformation grips, though any result from any landmark-based data analysis procedure can be so portrayed (see MacLeod, 2001; Zelditch et al., 2004).

Since its inception in the mid-1980s, geometric morphometrics has focused largely on the characterization of groups and statistical comparison of putative groups rather than the globally optimized group-discrimination problem per se. For example, of the 50 example and application chapters in the edited books *Contributions to Morphometrics* (Marcus et al., 1993) and *Advances in Morphometrics* (Marcus et al., 1996) only five include discussions of group-recognition issues and/or discriminant analysis results. When a discriminant analysis is pursued in the context of geometric morphometrics–either using the set of partial warp scores or the set of relative warp scores sometimes augmented by other variables (e.g. centroid size, scores on the uniform components of shape variation)–it is typically implemented through linear canonical variates analysis, usually with an accompanying MANOVA to test the statistical significance of group separation (see Baylac and Daufresne, 1996; Bogdanowicz and Owen, 1996; Demeter et al., 1996; Gubányi, 1996; Hugot and Baylac, 1996; Zelditch et al., 2004). This approach seems to work well enough for small data-sets, though there is an interesting inconsistency with the philosophical approach preferred by some geometric morphometricians (see Klingenberg and Montiero, 2005).

Artificial Neural Nets (ANNs)

Whereas two- and three-dimensional landmark-based morphometric approaches are well known and have a distinguished history of application in many systematic contexts, ANNs are just beginning to be understood by systematists and applied to systematic problems (see MacLeod, 2007c; MacLeod et al., 2007a,b), usually via analysis of sets of two-dimensional images. Artificial neural nets are able to make use of the three-dimensional information encoded in such images because they use the whole image or some down-sampled version thereof (see later discussion), rather than a set of purely geometric coordinate values marking landmark positions of tracing peripheral curves.

Artificial neural nets were developed originally as an alternative approach to computer design. What we now regard as typical 'computer programmes' are actually complex systems of instructions each of which has been devised by a human operator. These instructions are written in a symbolic code–the computer programming language that is, in turn, translated into a another symbolic language the computer system understands–the machine code–by a compiler, which is itself a computer programme. But the symbolic-instruction approach is not the only way to teach computers how to perform useful tasks. Just like humans, computers can learn by example. Instead of writing down complicated sets of instructions that specify desired outcomes for all possible actions that might take place in a complex process, one can arrange the computer programme to be nothing more than an interconnected set of simple switches (= processors) that store numerical values, one value per interprocessor pathway or neuron. Such an ANN can store information and devise its own rules for making appropriate responses to input data provided it is given user-mediated training as to what responses are desired.

This training process begins by constructing the ANN and assigning random numbers or weights to the various pathways. For example, say we have a series of digital sound files of instruments playing notes of the musical scale. The first note we submit to our network might be a harp playing A. The inputs to the network might be sensors tuned to different ranges in the audio spectrum. The harp A would be represented as a set of activation intensities from the various sensors. Since no two harps sound exactly the same, a series of harp As would be required to provide sufficient information about harp sound variation to allow the ANN to recognize an A note played by any harp. Network tuning would be accomplished by submitting a 'training' set of harp A-note recordings to the system and recursively adjusting the interneuron weights until all training-set signals (harp As) produced a clustered set of output signals. A similar process would be used to recognize A-notes for other instruments and then to teach the system to recognize other notes produced by these instruments. Eventually a neural net-based system could be assembled that would, in principle, be able to transcribe the musical notation for any passage that could be played or recognize an instrument type from its recorded sound. The advantages of this approach to computing include:

1. Lack of a necessity to devise specific algorithms and write computer programmes for complex operations (ANNs train themselves);
2. The same ANN architecture used on very different data types;
3. Flexibility in that properly designed ANNs can cope with substantial proportions of missing data–unlike morphometric approaches–and will work successfully on many holistic features that are difficult to characterize using simple descriptors;
4. Robustness to degradation (indeed, it can reprogramme itself if necessary);
5. The ability to get better at tasks the more such networks are used–to take advantage of 'on-the-job' training.

Both symbolic instruction and ANN approaches to computing have their strengths and weaknesses. The symbolic-instruction approach works well getting computers to perform simple, repetitive tasks quickly and accurately on data that can be represented by simple symbols–precisely the sorts of operations humans are bad at performing. Artificial neural nets often have the edge in performing complex, subtle tasks on generalized data, but they can be difficult to train and are heir to a variety of complexity-related problems (overtraining, non-deterministic performance, access to adequate training sets, convergence on suboptimal results,[8] etc.). Of course, these problematic aspects of ANN design are also characteristic of many complex algorithm-based systems.

In order to improve the performance of ANNs a number of variant designs have been developed, including the following (see Lang, 2007, and MacLeod et al., 2007a,b, for additional description/discussion):

Multilayer perceptrons (MLPs)–ANNs consisting of a fixed, layered structure that remains static during the training process. MLP nets are trained using supervised learning with weights adjusted during training iterations by a process termed backpropagation (see Bishop, 1996; Lang, 2007). Networks employing an MLP design are used routinely for generalized classification and identification tasks. While effective, MLP nets are time consuming to train and sample dependent (e.g. must be completely retained if a new group is added).

Kohonen self-organizing maps (SOMs)–ANNs in which the dimensions of the grid correspond to the number of input observations or variables and in which the network's structure is modified dynamically, based on the training-set data, until it forms an accurate representation of similarity relations among the different groups. In this design the nodes represent objects and the variables are used to estimate the interobject distances. As ANN training proceeds the internode distances are modified so that the structure comes to resemble a topographic surface reflecting phenetic similarity relations within the original data. To date SOM nets have been employed primarily to support data visualization, but they are becoming more common in a wide variety of applications. While Kohonen SOMs represent a better data-analysis strategy for understanding the relations between groups, they are more difficult to construct and are limited by their need to specify the size of a SOM at the outset of an analysis. This makes SOMs ideal for dealing with well-constrained problems (e.g. identification systems for well-known groups whose taxonomy is stable), but limits their overall applicability in open-ended situations.

Plastic self-organizing maps (PSOMs)–among the most advanced ANN designs, a PSOM net begins with a few nodes connected by random links. When data from an initial group are presented to the network more nodes are created and the existing internode weights adjusted so that they 'move' closer to the established group. As more examples of the initial group are added, high-valued internode weights are further strengthened and low-valued weights diminished, changing the structure of the ANN. When data from a new training group are introduced, a new set of nodes is created and placed in an appropriate position relative to the established topology. As more data from the new group are evaluated, the weights of high-valued internodes are adjusted, eventually to the point where most or all internodal connections between established groups are lost. This process is then repeated until all groups are located relative to one another in the dynamic network topology. PSOM nets have yet to be applied widely, but bench-top trials on small data-sets suggest they hold promise for being able to address the scalability problem.

A COMPARATIVE ANALYSIS

Now that the basic concepts of geometric morphometrics and ANNs have been reviewed it is time to move to a detailed consideration of strategies, and evaluation of results generated by using these two approaches to analysing shape data. To facilitate direct comparison, an example data-set will be used consisting of two-dimensional images and three-dimensional surface scans of the left valve of three species belonging to the bivalve genus *Astarte* in plan view.

MATERIALS AND METHODS

MATERIALS

Astarte is a small to medium sized (<7 cm), infaunal, heterodont bivalve characterized by an equivalve, inequilateral lateral shell composed of thick, trigonal to trigonally rounded valves characterized by deep and well-defined lunules and escutcheons along with prominent prosogyrate beaks (Todd, 2000; Figure 9.7). The genus first appears in the fossil record during the Jurassic (Bajocian, *c.*170 Ma) and continues to be represented by extant species. Both subgenera and species are recognized by morphological criteria, especially structural details of the gross valve geometry, hinge, lunule, escutcheon and especially the size, shape, frequency and distribution of the pronounced comarginal rugae/ribs.

Specimens of three fossil astartid species were selected for this analysis: *A. mutabilis*, *A. obliquata* and *A. omalii* (Figure 9.8; Plates 9.1–9.3). These specimens were all obtained from sorted collections of material from Red Crag (Upper Pleistocene) and Coraline Crag (Middle Pleistocene) in the mollusk collections

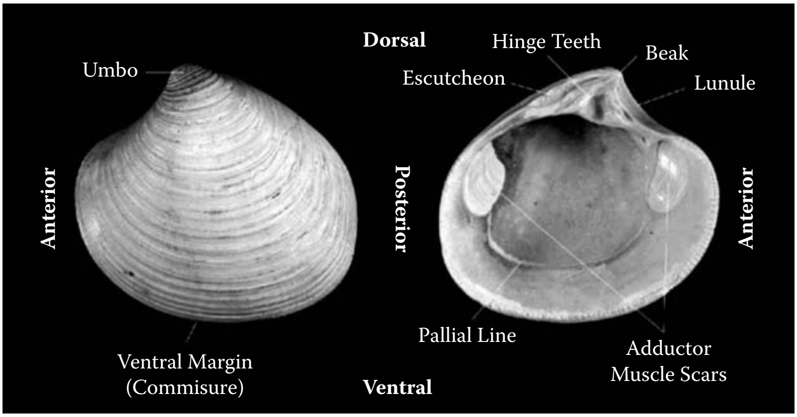


FIGURE 9.7 Left valve of *Astarte donax* illustrating the primary morphological features of the bivalve shell.

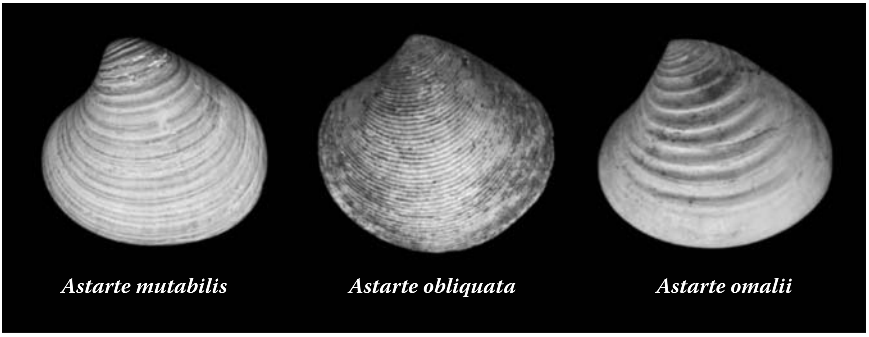


FIGURE 9.8 Examples of the three *Astarte* species used in the example analysis.

of the Natural History Museum (London). *Astarte mutabilis* is characterized by a thick, vaulted, transversely oblong, inequilateral, heart-shaped shell with broadly sulcated, curving umbones and smooth (young specimens) to thick and deeply crenulated (mature specimens) margins (Figure 9.8; Plate 9.1). In contrast, *A. obliquata* is typified by a thin, flattened, ovate, slightly inequilateral shell with prominent, sharp umbones. Both *A. obliquata* valves are also densely ornamented with closely packed, broad, flat-topped, oblique striae (Figure 9.8; Plate 9.2). Finally, *A. omalii* is, in many (but not all) ways, intermediate between *A. mutabilis* and *A. obliquata*. Wood (1850–1856) reports *A. omalii* as being ovately oblong to trigonal, smooth to sulcated, tumid or compressed, for the most part inequilateral, with sharp, sulcated umbones and a crenulated valve surface, while noting that in some mature specimens the surface is smooth (Figure 9.8; Plate 9.3).

The broadly qualitative nature of the foregoing descriptions is, of course, typical of much of the systematic literature. Professional practitioners, as well as students, have traditionally been left to develop their own concepts of the limits of species variation, usually from inspecting either plates of drawings or photographs (whose number is inherently limited), museum collections (in which the species in question may or may not be present and, if present, are usually represented by a small number of specimens) or collecting localities (which may or may not still exist). Fortunately, in the case of these three species it was possible to assemble small reference collections of 30 specimens each on which to base morphometric analyses. As mentioned previously, these specimens all came from one of two well-known UK Pleistocene localities and were likely identified by Wood during the study that resulted in his 1850–1856 monograph, *Mollusca from the Crag.*

Data Collection Methods

Digital Photography

The left valves of all specimens were oriented on a copy stand such that the plane defined by the valve's margin or commissure was normal to the camera's optic axis. Specimens were then digitally photographed using a standardized lighting configuration against a black background at a resolution of 150 dpi and saved as uncompressed

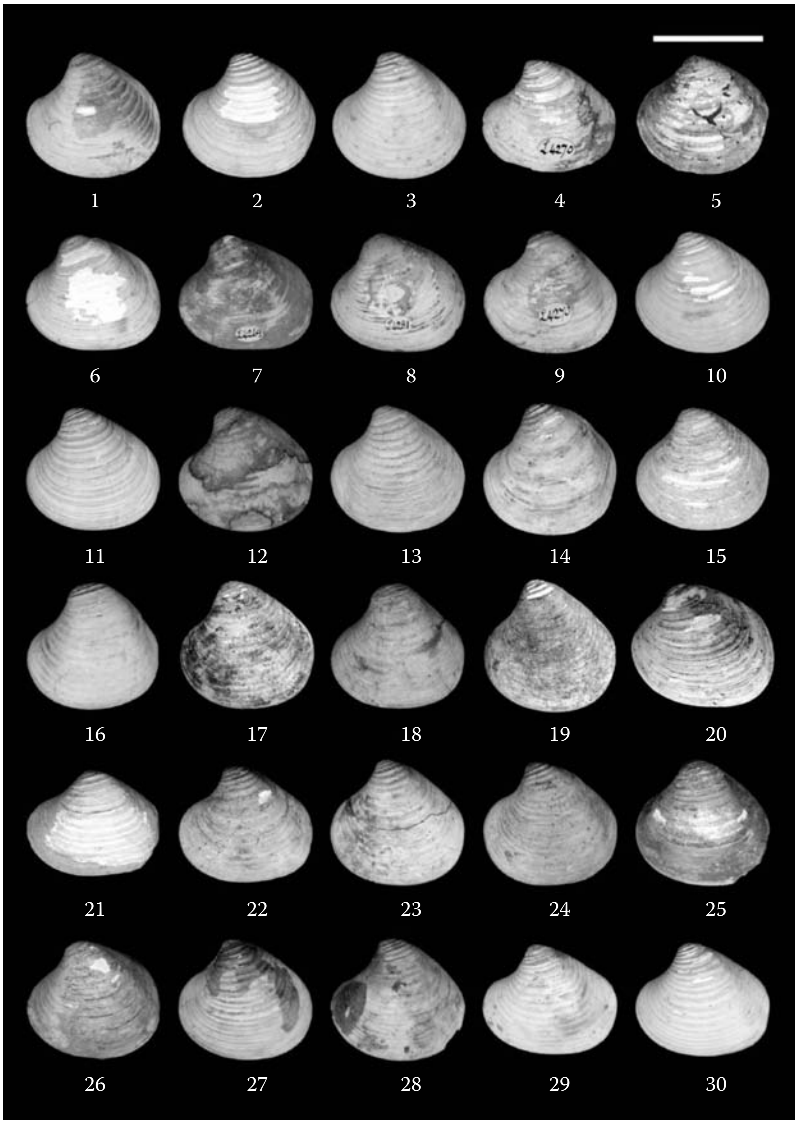


PLATE 9.1 The *Astarte mutabilis* dataset.. Value of the scale bar for each image as follows (in cm): 1. 2.44; 2. 2.78; 3. 2.26; 4. 2.97; 5. 4.05; 6. 1.85; 7. 1.93; 8. 1.89; 9. 2.30; 10. 1.85; 11. 2.49; 12. 2.76; 13. 2.59; 14. 3.00; 15. 2.78; 16. 3.37; 17. 2.53; 18. 2.64; 19, 3.16; 20. 1.97; 21. 3.48; 22. 2.69; 23. 2.26; 24. 2.28; 25. 2.35; 26. 2.18; 27. 2.44; 28. 2.31; 29. 2.48; 30. 2.09.

tiff files. Image postprocessing consisted of cropping to a standardized size, correcting the exposure to achieve a high-contrast, well-balanced image, editing out any background artifacts (e.g. dust particles, background glare), and employing a single-pass unsharp mask (50%) to bring out surface detail. See Plates 9.1–9.3 for results. These two-dimensional digital photographs were used to document the morphology expressed by each specimen in each of the three data-sets, to serve as a reference

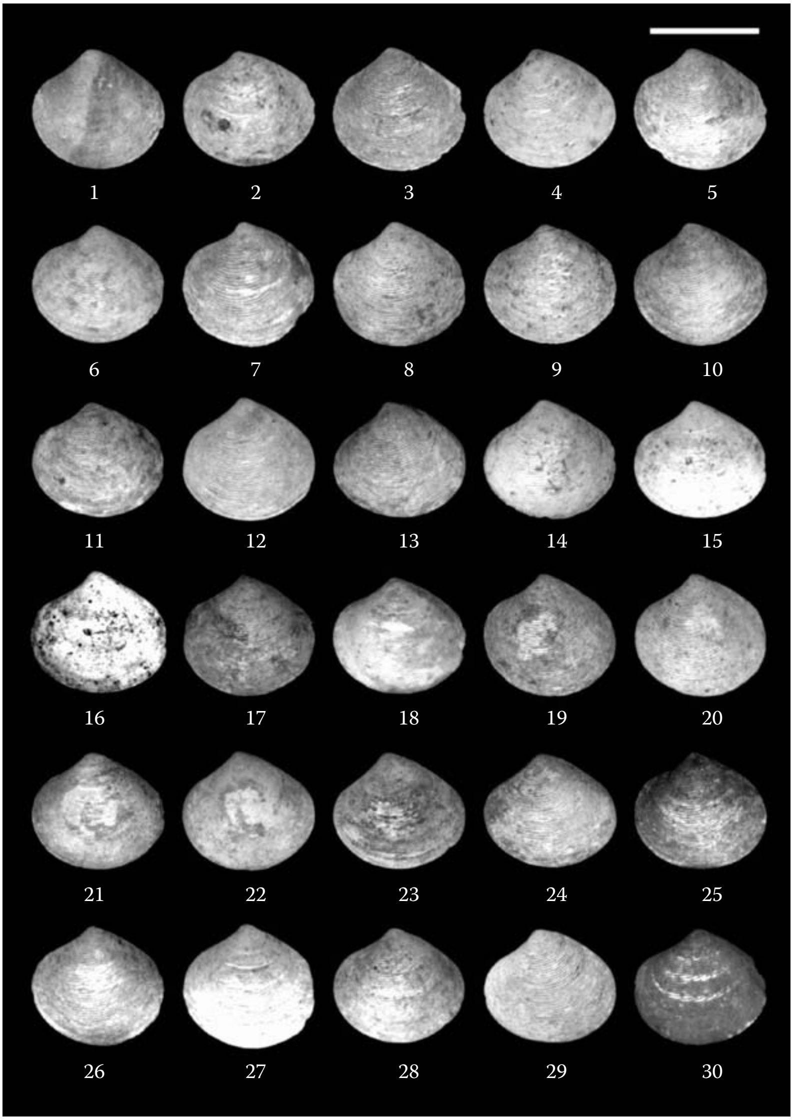


PLATE 9.2 The *Astarte obliquata* dataset.. Value of the scale bar for each image as follows (in cm): 1. 3.22; 2. 3.96; 3. 4.40; 4. 3.87; 5. 4.63; 6. 3.11; 7. 4.63; 8. 3.82; 9. 3.32; 10. 3.56; 11. 4.30; 12. 4.25; 13. 4.05; 14. 4.10; 15. 4.58; 16. 4.54; 17. 3.31; 18. 3.35; 19. 3.96; 20. 4.83; 21. 3.29; 22. 3.48; 23. 3.96; 24. 3.63; 25. 4.14; 26. 4.00; 27. 4.41; 28. 3.96; 29. 3.27; 30. 4.41.

for evaluation of data derived from the three-dimensional scans and as input to the ANN analyses.

Three-Dimensional Laser Scanning

The left valves of all specimens were oriented on a mounting board such that the plane defined by the valve's margin or commissure was normal to the scanner's optic

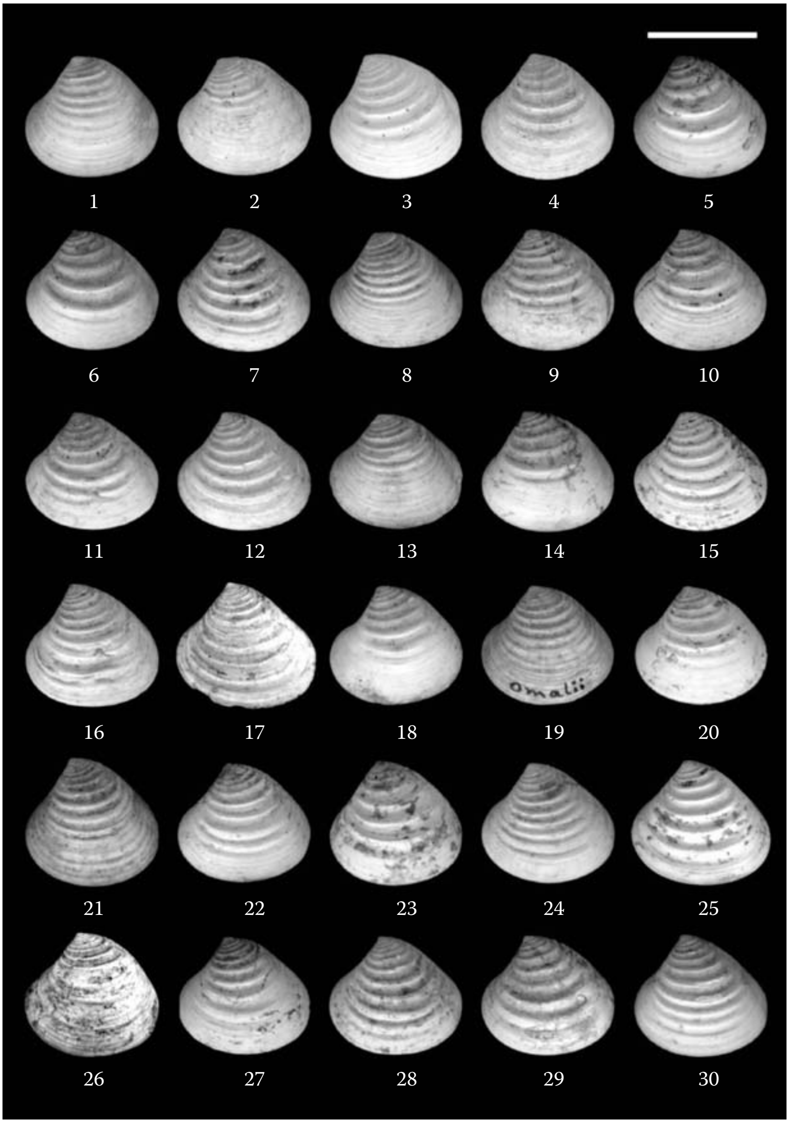


PLATE 9.3 The *Astarte omalii* dataset.. Value of the scale bar for each image as follows (in cm): 1. 3.00; 2. 2.64; 3. 3.00; 4. 3.14; 5. 3.22; 6. 3.26; 7. 3.35; 8. 2.88; 9. 3.11; 10. 3.20; 11. 3.38; 12. 3.00; 13. 2.67; 14. 3.22; 15. 3.28; 16. 3.63; 17. 3.53; 18. 3.45; 19. 2.92; 20. 3.96; 21. 3.16; 22. 3.35; 23. 3.67; 24. 2.90; 25. 3.24; 26. 2.91; 27. 3.31; 28. 3.22; 29. 3.11; 30. 3.26.

axis. The valves were then scanned using a Konica-Minolta *VIVID 910* laser scanner (spatial resolution *c*.0.01 mm). Because details of the dorsal area are obscured in plan view, secondary scans were performed with the specimens oriented such that the scanner was focused on the dorsal area. Once these oblique scans had been obtained the two scans for each specimen were composited using the three-point method outlined previously. Scan postprocessing involved cropping out scan artefacts, cutting the scans above the line of the lunule and escutcheon (so that the scan surfaces would be single valued), filling holes in the scan mesh, decimating the scans from their original resolution (*c.*20,000–40,000 vertices) to *c.*5000 vertices and adaptively remeshing the scans so that regions with more surface variation would be represented by a greater number of vertices. These steps are illustrated in Figure 9.2. The resultant three-dimensional scans represent virtual facsimiles of the left valve's lateral surface that efficiently represent morphology of that surface at an average resolution of *c.*0.25 mm. These three-dimensional scans were then exported in a variety of forms and formats to supply data for various stages of the analysis:

1. Entire scan exported as a standard tessellation language (STL) file for archive purposes;
2. Entire scan exported as an ASCII *x,y,z* point file for eigensurface analysis (see later discussion);
3. Scanned surface boundary points exported as an ASCII *x,y,z* point file for eigensurface analysis (see later discussion);
4. Coordinates of two semilandmark points on the surface boundary export as an ASCII *x,y,z* point file for eigensurface analysis (see later discussion). These points were the valve beak and the midpoint of the distal shell margin in lateral or plan view; and
5. An image of the virtual model oriented in the same manner as the original scan (see earlier discussion) exported as a tiff file for use in the PSOM-ANN analysis (see Plates 9.4–9.6).

Data Analysis Methods

In order to compare the performance of geometric morphometric approaches to group characterization, a number of numerically equivalent analyses were performed.

Three-Dimensional Relative Warps Analysis

For the *Astarte* data-set four landmarks were chosen as the minimum necessary to quantify the gross shape of the valves. These were the tip of the beak, midpoint of the ventral commissure and extremal points of the right and left valve margins when the shell is oriented in standard position (the line between the centres of the adductor muscle scars parallel with the page margin; see Figure 9.9). The first of these is a Type 2 (extremal point, or maximum of curvature) landmark in the Bookstein (1990) classification. All the others are Type 3 landmarks (semilandmarks located via geometric reference to other landmarks). Together they capture an estimate of the size and general shape of the shell in the manner of a four-point polygon.

Such data quantify aspects of each valve's obliquity–which is a fundamental taxonomic attribute of the difference between these species–but little else. It is also of interest to note that (1) while these landmark points are not coplanar, they are nearly so and thus provide no information about the relative vaulting of the coiled shell and (2) all landmarks are located on the valve's outline. In bivalves, as it is for most biological forms, outline shape is of fundamental importance to the taxonomy,

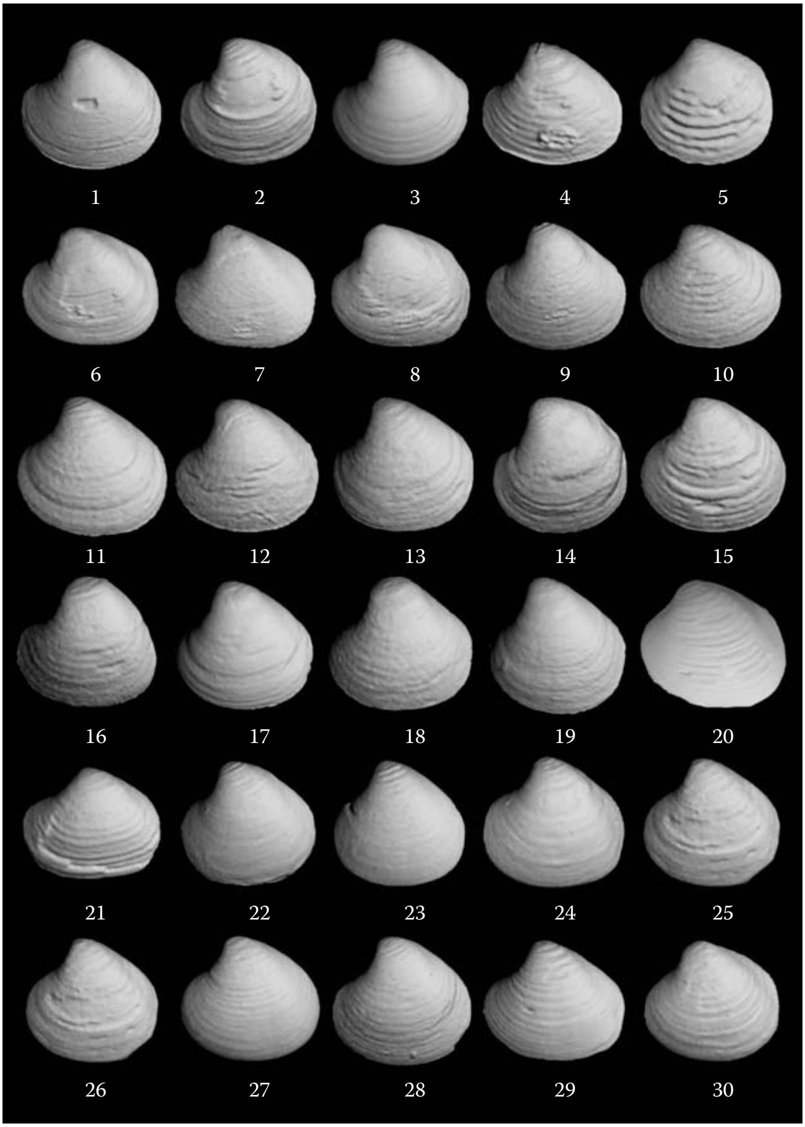


PLATE 9.4 Adaptively meshed virtual models of the *Astarte mutabilis* dataset (see Plate 9.1). These *x,y,z* data were used as input to into the 3D eigenshape and eigensurface analyses and these images were used as input into the *DAISY* PSOM-ANN analysis.

life history and functional range of the organism. It is, therefore, no coincidence that, despite the actively antagonistic stance taken by many 'landmark-centric' theorists and practitioners (e.g. Bookstein et al., 1982; Bookstein, 1990, 1991; Zelditch et al., 1995, 2004), most applied morphometric investigations employ landmark configurations that are dominated by semilandmark points that lie on the specimen's outline (e.g. see example analyses in Bookstein, 1990, 1991; Marcus et al., 1993, 1996; Zelditch et al., 2004).

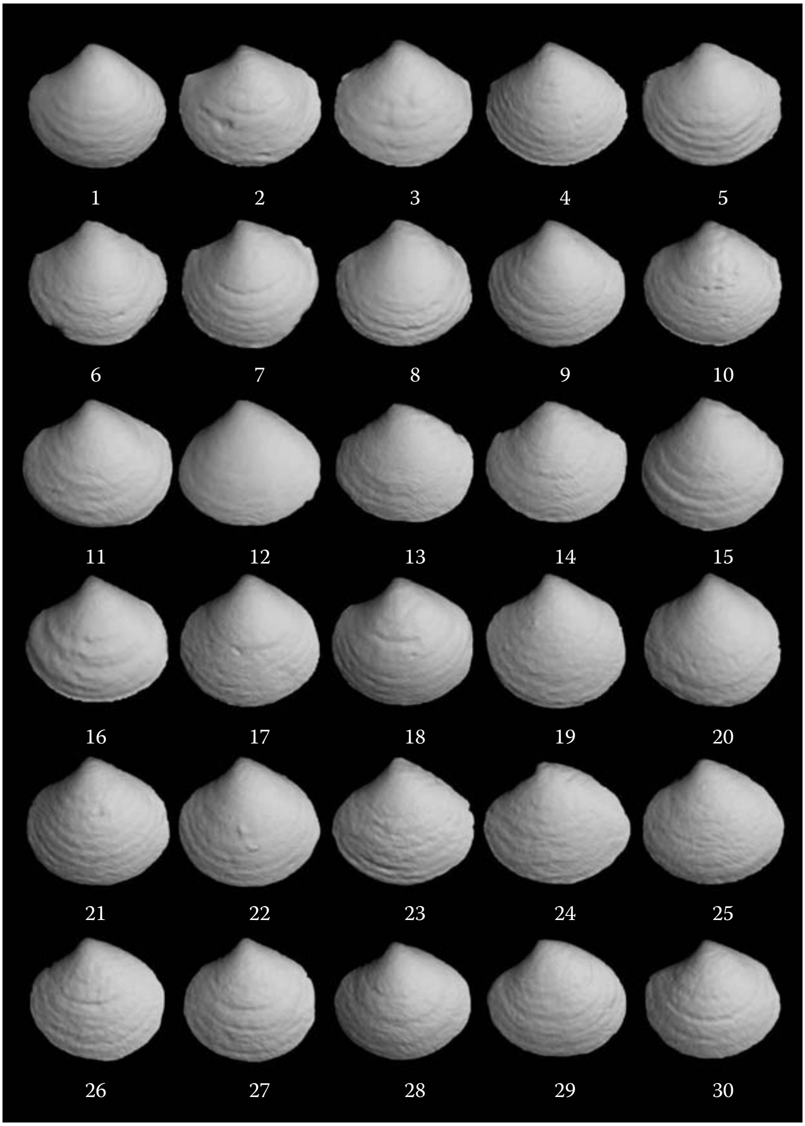


PLATE 9.5 Adaptively meshed virtual models of the *Astarte obliquata* dataset (see Plate 9.2). These *x,y,z* data were used as input to into the 3D eigenshape and eigensurface analyses and these images were used as input into the *DAISY* PSOM-ANN analysis.

Landmark data were collected from the three-dimensional scan files in the form of sets of three-dimensional point coordinates. These data-sets were used to quantify the size of the valves using the centroid size statistic (Bookstein, 1986) and then translated, rotated and scaled using the (GLS) superposition method (Rohlf and Slice, 1990). Once the size-free Procrustes shape coordinates had been obtained in this manner each specimen's shape was represented by a shape 'function' (or, more correctly, a vector) consisting of the ordered shape coordinate values (Figure 9.9).

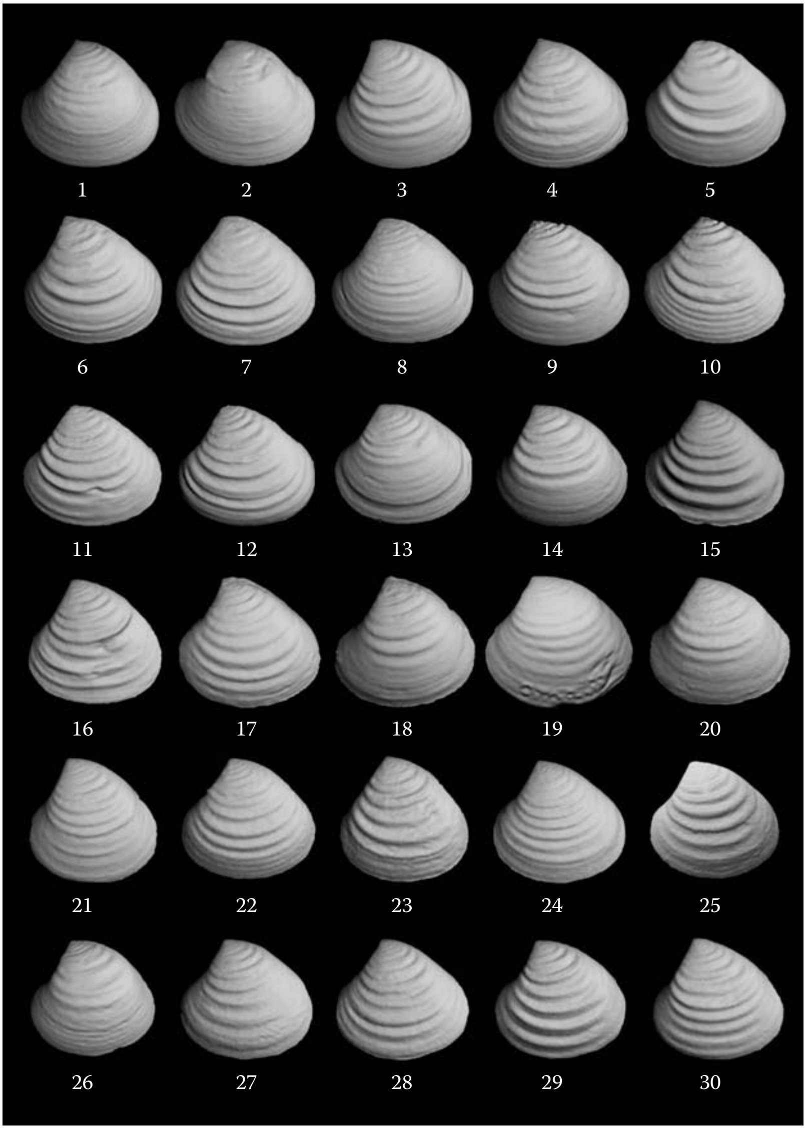


PLATE 9.6 Adaptively meshed virtual models of the *Astarte omalii* dataset (see Plate 9.3). These *x,y,z* data were used as input to into the 3D eigenshape and eigensurface analyses and these images were used as input into the *DAISY* PSOM-ANN analysis.

Geometric similarities in the form of these shape functions/vectors represent corresponding similarities in the geometry of the valves. Shape similarities and differences were summarized by calculating the pair-wise covariance between shape functions by decomposing this matrix hierarchically using singular value decomposition (SVD). For square, symmetric matrices this operation results in specification of a series of mutually orthogonal singular axes that are equivalent to major axis multivariate regressions through the set of shape vectors representing the sample.

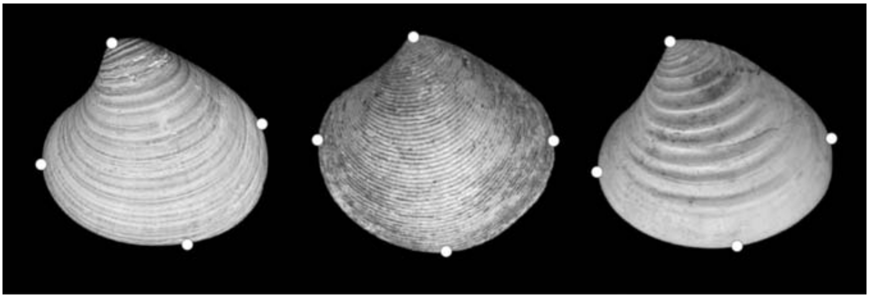


FIGURE 9.9 Landmarks used to quantify astartid valve shape for the three-dimensional relative warp data analysis. Species as in Figure 9.8.

Coordinate positions along these vectors represent a trajectory of hypothetical shapes that account for linear aspects or modes of the shape variation present within the sample. Their relative lengths express hierarchical dominance of these modes subject to the mutual orthogonality constraint. All operations described to this point are formally equivalent to what has been termed 'relative warps analysis' by Bookstein (1990, 1991), Rohlf (1993) and Zelditch et al. (2004), though in this discussion I have not emphasized their relation to the geometry of the Kendall shape space or the thin-plate spline.

These sets of sample-specific, hierarchically arranged latent shape vectors were then used as mutually independent shape descriptors for re-expressing the observed shape variation in a lower dimensional space that optimized the shape signal-to-noise ratio. A cutoff value of 95 per cent of observed shape variation was used to determine how many summary vectors were needed to adequately characterize the sample. Once the original data had been re-expressed by projection onto these vectors, the new, shape variance-optimized data were subjected to canonical variates analysis (CVA; see MacLeod, 2007b, and references therein) to define a linear space where differences between the a priori designated species groups could be optimized, portrayed and analysed. While this operation destroys the inherent geometry of the Procrustes shape space (Klingenberg and Montiero, 2006), the purpose here is not to assess the geometry of that shape space, but rather to summarize and assess the limits of group-specific shape variation as well as to create an optimized space that can be used to assign unknown objects to the correct species group with confidence.

Three-Dimensional Eigenshape Analysis

A modified form of extended eigenshape analysis (MacLeod, 1999, 2002, 2005) was used to compare and contrast the results that could be gained by focusing on specimen outlines as opposed to just a few landmarks. Eigenshape analysis begins by interpolating coordinate-point data collected from an outline into a series of landmarks (usually the starting point for outline digitization) and semilandmarks arranged such that distances between adjacent landmarks is constant (Figure 9.10). If landmarks in addition to the starting point can be located on the outline, these can be used to subdivide the outline into corresponding segments.

Because the rationale for using semilandmarks to represent a specimen's outline has often been (e.g. Bookstein et al., 1982; Bookstein, 1990, 1991; Zelditch et al.,

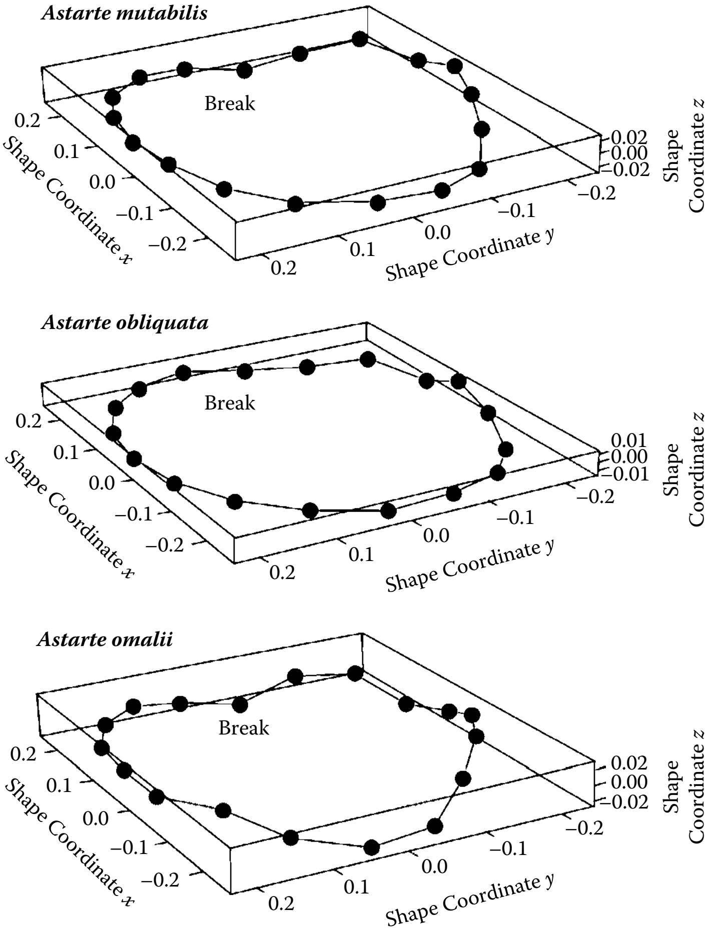


FIGURE 9.10 Perspective scatterplots showing representative three-dimensional outline data for the three example astartid species. These outlines have been transformed to Procrustes shape coordinates in order to remove the effects of size, positional and rotational differences. Note degree of relief along the morphometric *z*-coordinate axis, which is low overall but does differ strongly among these three specimens.

1995, 2004) and because the eigenshape sampling strategy will be extended to three-dimensional surfaces in the following section, it is worth taking a moment to review the difference between strictly landmark-based and outline-based morphometric analyses. As noted previously, landmarks represent positions that can be located unambiguously on each shape in the sample. Ideally, these represent the juxtaposition of different structures that define a unique point (Bookstein's Type 1 landmark class, e.g. the intersection of the leading margin of a fish's fin with the body, intersection of three bones of a vertebrate skull). More often they are arbitrarily designated positions such as the centres of structures or their extremal points (Bookstein's Type 2 landmark class, such as the centre of the eye, tip of a fish's fin, maximum of curvature of a skull bone in dorsal view) or semilandmarks. The important things to note about these eminently practical ways of defining landmark positions are that (1) they represent useful guides for morphological comparison regardless of the fact that such point locations can rarely (if ever) be construed to meet the formal definition of biologically homologous structures (see MacLeod, 1999, 2002a; Humphries, 2002), (2) the purpose of these landmarks is to locate the structure or some part thereof relative to other structures or parts of structures within the form and (3) such designations have direct conceptual links to the qualitative descriptions of shape used routinely by taxonomists/systematists to both define and discuss the characteristic form of taxa, limits of variation allowable under particular species concepts, etc.

When extending this geometric formalism to the consideration of outlines, the appropriate linking concept is not one that equates landmark locations with the placement of semilandmarks along an outline or outline segment. Rather, the most appropriate link is between a landmark (or multilandmark landmark set) specifying the position (and shape) of a structure and the *entire* outline or outline segment represented by the entire collection of semilandmarks that quantify the structure's position and shape. Both these collections of geometric data represent the positions and shapes of corresponding structures relative to other such structures, but neither the landmark nor semilandmark sets can be considered biological homologues to other landmarks or semilandmark sets at the level of individual point locations. The reason for this is that the concept of biological homology refers to entire structures, not point locations on structures, and can only be recognized in the context of an explicit phylogenetic hypothesis (see MacLeod, 1999). Such point locations do embody the concept of topological correspondence, sometimes confusingly referred to as 'topological homology'.

Additionally, it is important to remember that the integrity of the geometric description of the structure's position and shape is literally the point of the exercise. The important task of maintaining topological correspondence cannot be ensured if landmarks do not remain conceptually and mathematically tied to the relative positions of the structures they represent, or if landmark sets, outlines or outline segments do not remain conceptually and mathematically tied to both the relative positions *and the shapes* of the structures they represent. Of course, it should also be appreciated that outlines add substantially more relevant biological information to a morphometric analysis than small sets of landmarks can ever hope to contribute precisely because they sample more of the relevant geometrical information (see MacLeod and Rose, 1993; MacLeod, 1999, 2002; Adams et al., 2004; MacLeod, 2004). Certainly the inclusion of outline-based information will enable morphometric analyses to better test qualitatively formulated systematic hypotheses that refer to outline-based morphological characteristics.

With these concepts, correspondences, and differences in mind the three-dimensional shapes of the *Astarte* shells were quantified by using landmarks 1 and 2 (see previous discussion) to divide the boundary outline into two corresponding segments: an anterior half-outline and a posterior half-outline. The outline curves for each segment from each adaptively meshed virtual specimen were then interpolated to eight equally spaced semilandmark points (Figure 9.10). This sampling scheme ensures that neither the posterior nor anterior curve segment was differentially weighted in the analysis in the sense of being represented by proportionately more data. In terms of fidelity to the original shape, this reduced resolution induces a less than five per cent error to the character of the original curve as represented by the remeshed virtual specimens across the entire sample.

In a departure from standard eigenshape procedure (see MacLeod, 1999) the interpolated three-dimensional outline semilandmark coordinates were not transformed to the Zahn and Roskies (1972) angular (ϕ) shape function, but were subjected to standard Procrustes (GLS) superposition (see also MacLeod, 1999). While the Procrustes superposition approach is not mathematically equivalent to the ϕ function, both methods accomplish the same translation, rotation and scaling transformation and produce similar results, though one that is less compact in the case of the Procrustes method. Nevertheless, since neither the three-dimensional landmark nor eigensurface data could be represented by ϕ shape functions, it was felt this departure from standard practice was warranted to facilitate precise comparisons among the various results.

Once the specimen outlines had been superposed, the shape-coordinate data were subjected to an SVD analysis to define the set of independent shape axes that accounted for 95 per cent of the observed shape variation. Scores on the original three-dimensional Procrustes shape vectors on these axes were then used as input to a CVA. Results enable an examination of the extent to which these three-dimensional outline data recover distinctions among the three *Astarte* species and define a discriminant space that could be useful in identifying unknown specimens.

Eigensurface Analysis

Obviously, both landmark and outline approaches focus on only a small subset of the morphological data available to the systematist for qualitative analyses. Until recently, little progress had been made in extending morphometric analysis to handle the analysis of three-dimensional surfaces. For the most part this has been due to difficulties in obtaining three-dimensional surface data in the form of discrete *x,y,z* coordinate points. However, over the last few years, accurate and relatively inexpensive laser scanners have begun to appear on the market. While primarily used for computer-aided design (CAD), modelling and reverse engineering applications, such devices have the potential to collect data in forms useful to morphometricians and systematists. Eigensurface analysis was–and continues to be–developed as one approach to the use of such data (Polly and MacLeod, 2006, MacLeod and Polly, 2005). The algorithm and approach described next represent the latest version of eigensurface analysis and incorporate a number of features not available in previous versions.

Eigensurface analysis is a more or less direct extrapolation of eigenshape analysis as described in the preceding section to the analysis of three-dimensional surfaces. In the same way that eigenshape analysis fits an equal series of regularly spaced points to an outline or outline segment in all specimens across a sample in order to quantify the shape of the various curves, eigensurface analysis fits an equal series of regularly spaced points in the form of a grid to the surface of all specimens across a sample in order to quantify the shape of the various surfaces. Like eigenshape analysis, this operation is accomplished through location of a series of semilandmark points and tied to biological reality through user-specified landmark points that ensure the grid nodes represent the same topological positions across all specimens. Also like all eigenshape methods, the basis of comparison within eigensurface analysis is the complete set of grid points that represent the measured surfaces to a specified resolution or tolerance level.

All landmark points (true landmarks and semilandmarks) participate in the analysis in exactly the same way because all are needed to represent the surface being analysed. In terms of the current application, eigensurface analysis makes use of four sets of information (Figure 9.11). The first of these is the cloud of *x,y,z* coordinate points placed on the surface of the object by the scanner. The second is the set of *x,y,z* coordinate points located, in this case, on the periphery or outline of the shell (in the case of complete virtual specimens this would be an outline located along a suitable trace that would subdivide the specimen into meaningful sections–for example, dorsal–ventral, right–left, interior–exterior). The third is a set of two landmark points that are used to control placement of the grid on the specimen in a manner similar to extended eigenshape analysis (MacLeod, 1999, 2002a, 2005; Barrow and MacLeod, in press). The fourth is a grid resolution (e.g. 10-point, 20-point, 50-point).

The eigensurface algorithm begins with a PCA transform of the *x,y,z* coordinate point cloud, outline and landmark data-sets to bring the orientation of the virtual objects into an approximate alignment. The landmark points are then used to subdivide the outline into two sections (Figure 9.11, upper row). Interpolation is used to place a series of equally spaced points on each of the two outline segments. The number of points used to represent the points is equal and constrained to match the dimension of the sampling grid. Since the landmark points will not necessarily divide the specimen into two symmetrical halves, the intersemilandmark distances will be constant within an outline segment, but differ between the two outline segments. Accordingly, it is important to select a grid resolution that will represent the geometry of the outline and the surface adequately (unless morphological sampling issues are a target of the analysis; see later discussion).

Once the outline has been sampled, a chord joining the two landmark points is projected onto the surface of the virtual object (represented by the point cloud data, Figure 9.11, upper row) using a nearest neighbour approach to implement the *z*-coordinate mapping. This chord represents the ‘backbone’ of the grid and ensures the grid’s orientation is consistent and matched to topologically corresponding locations. In principle, any number of landmark points could be used to place the grid backbone onto the surface of the object and ensure that the location of topologically corresponding substructures on the specimen’s surface were always located at the same relative grid location. However, since the surfaces of these bivalve shells do not exhibit recognizable substructures, a straight chord was used for the backbone.

After the backbone chord had been located, three-dimensional interpolation was used to subdivide the chord into an equally spaced series of segments with the number of interpolated points being equal to the grid resolution and the number of points located previously along the outline segments. The purpose of the backbone chord is to help the grid adapt to the surface morphology of each shape and represent that shape by the greatest number of evenly spaced points consistent with the specified

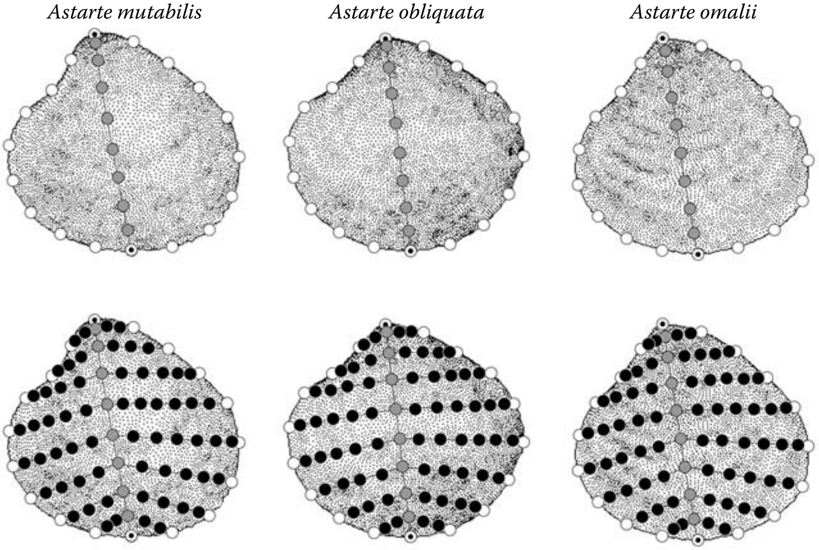


FIGURE 9.11 Steps in the eigensurface analysis grid-definition procedure. This begins with a cloud of *c*.5000 *x,y,z* points representing the valve's surface morphology. From these data, two landmarks (⊙ symbols) are selected that define the chord from the tip of the beak to the centre of the ventral margin (upper row). This chord is then projected onto the surface points using a nearest neighbour algorithm. The interlandmark chord forms the 'backbone' of the adaptive grid system. Simple interpolation is then used to find the positions of a series of semilandmark points along this backbone chord and each of the valve half-outlines defined by the original landmark points (● and ○ symbols respectively, upper row). The same number of points is interpolated among all curve segments, which also share starting and ending points. The number of points is set by the user and determines the resolution of the surface sampling grid. In this example a 10-point grid is illustrated (the outline segments and backbone are each represented by 10 points). Once the backbone and outline segments have been interpolated, the grid is completed by sampling along a series of grid-rib chords connecting the backbone and outline nodes to one another (lower row). These form chords whose trace is projected along the surface with semilandmark points (● symbols) interpolated such that spacing equal to or less than that of the interpoint spacing along the backbone chord is maintained. Each specimen in the sample is gridded in the same manner. The number of corresponding grid-rib nodes are then compared across the sample and the final dimensions for each rib set using the maximum number of rib nodes necessary to represent the form of the most shape-rich corresponding rib in the sample. The grid-rib chords are then re-interpolated from the original surface point cloud data to consistent numbers of intrarib nodes (lower row). This procedure can be applied to any sample of three-dimensional surfaces and results in all surfaces being represented by a set number of topologically corresponding semilandmark points.

grid resolution. Rectilinear grids often badly under-sample morphology in regions near the termini of the grid axes (Figure 9.12). This is especially problematic as there is often a tendency to orient the sampling grid to begin and end in regions that are relatively well constrained by biological landmarks–precisely the regions in which both topological and biological information is concentrated. Ideally, such regions should be well represented by sampling grid points. By constructing a sampling grid

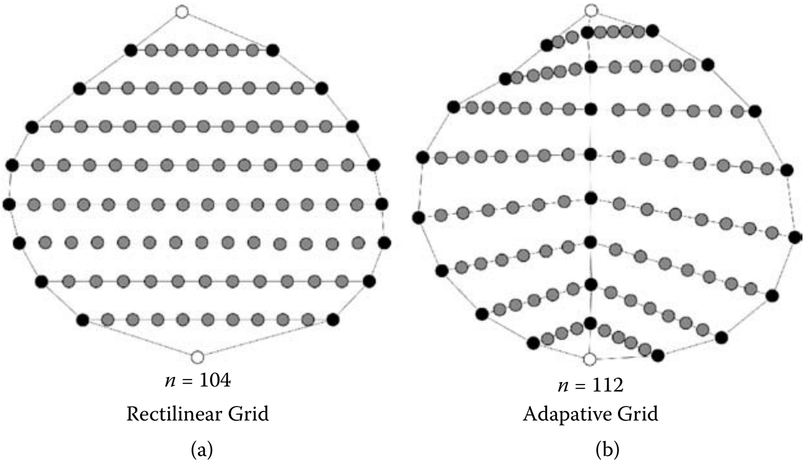


FIGURE 9.12 Alternative approaches to creating a grid of points to define a surface. Under rectilinear gridding (a) landmark points (white symbols) are used to define a chord that can be interpolated to a series of equally spaced segments. These segments define the grid. The points at which the segments cross the specimen's outline (black symbols) can then be used to define chord segments, which can themselves be interpolated to sets of equally spaced nodes (grey symbols). This procedure results in a regular grid, but leaves relatively large portions of the shape near the landmark points undersampled. Compare this with the result of a 10-point adaptive grid procedure (b; see Figure 9.11 and text for an explanation of the adaptive gridding procedure).

out of links between the outline and backbone vertices, this 'edge effect' is reduced. Moreover, use of the backbone chord to control grid placement provides a means by which the user can design the grid geometry to optimize the aspect of the surface sampling for the problem under consideration.

After an appropriate number of points have been specified along the outline and backbone segments, the sampling grid is specified by linking the backbone segment nodes to the outline segment nodes to form straight chords (Figure 9.11, lower row). These chords are then projected onto the surface of the virtual object in a manner identical to that used for the backbone chord, with the number of nodal points required to achieve an even internode spacing set to equal that used for the backbone node spacing. Of course, this grid-rib internode spacing will differ for each specimen in the sample due to differences in the surface shape. After the appropriate grid-rib internode spacing has been determined for each specimen, the complete set of grid-rib point resolutions is scanned and a sample-specific consensus resolution determined for each rib by selecting the largest value for each corresponding grid rib across the entire sample.[9] This procedure is time consuming, but ensures production of a grid resolution that has been rendered consistent across the entire sample.

Once these sampling grid parameters have been specified, all that remains is to resample each of the grid-rib surface-curve segments for each specimen using the set of consensus grid-rib resolutions. The surfaces for each specimen are then represented by grids of identical sizes and shapes that have been adaptively fit to each virtual surface and/or object. The set of grid nodes represents semilandmark point locations that are evenly, though not equally, placed across the surface and specify a set of topologically homologous locations. Because the point of eigensurface analysis is to compare actual surfaces with one another, no sliding of these semilandmarks in order to achieve maximum internode topological correspondence by minimizing some mathematical (but biologically irrelevant) parameter such as 'bending energy' was allowed (see Discussion section).

After three-dimensional grids have been obtained for each specimen, the surface can be thought of as being represented by a high-dimensional shape function or multivariate vector. The set of shape functions for a sample can then be registered using the three-dimensional Procrustes GLS superposition represented as a set of deviations from the mean shape (Figure 9.13) and the resulting shape-variable representation of the surfaces subjected to a covariance-based SVD. As mentioned previously, this procedure is mathematically equivalent to a standard relative warps analysis.

Results of the SVD analysis were then used to analyse (1) the ordination of surface shape in a low-dimensional linear space tangent to the Kendall shape manifold for these surfaces at the mean shape, (2) modes of shape variation represented by the analysis (via examination of the eigenvalues and use of the singular vectors [= eigensurfaces] to create virtual models of shape variation) and/or (3) as a geometric transformation that optimizes the shape signal-to-noise ratio. When the latter was employed the scores (= covariances) of the sampled shapes with the first *n* eigensurfaces whose combined squared singular values (= variances) add up to a user-defined proportion (e.g. 95%) of the total shape variance exhibited by the sample were used to achieve this optimization. These scores were also plotted to achieve a direct visualization of

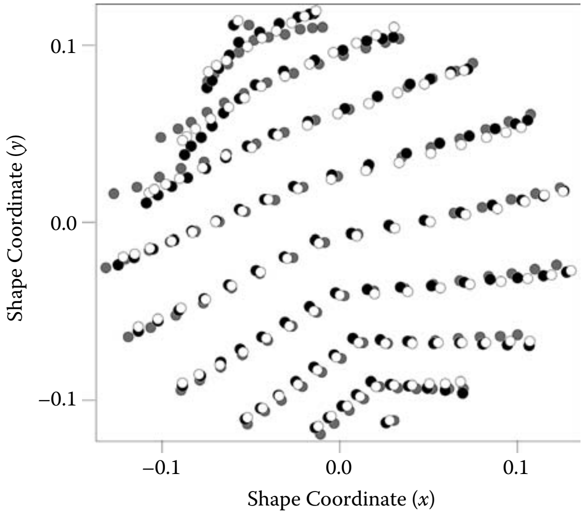


FIGURE 9.13 Procrustes (GLS) superimposition of 10-point eigensurface grids for the three *Astarte* specimens shown in Figure 9.8: *A. mutabilis* (black symbols), *A. obliquata* (grey symbols), *A. omalii* (white symbols). Note discreteness of most semilandmark-specific point clouds other than in the areas of substantial interspecific shape change (beak and umbo, anterior and posterior regions). See text for discussion.

surface shape similarities and differences present in the sample and used as input to statistical tests and other analytical procedures (e.g. CVA).

PSOM ANN Analysis

Results from application of the morphometric approaches described before were compared and contrasted with results obtained from application of PSOM net analyses of images of both actual specimens and virtual representations of specimens in order to assess this alternative approach to characterizing taxonomic groups quantitatively. The PSOM approach to species identification was implemented by the digital automated identification system (*DAISY*; see Weeks et al., 1997, 1999a,b). *DAISY* was chosen as the best currently available implementation of a PSOM net that has been programmed specifically for use in systematic applications (see MacLeod, 2007c).

The *DAISY* system accepts training sets in the form of standard-format images (e.g. tiff) of authoritatively identified specimens. Each image was processed by reducing its spatial resolution (via subsampling) to a 32 × 32 pixel grid, adjusting its pixel-level spectrum to achieve brightness equalization and transforming each pixel grid from a Cartesian to a polar format (Figure 9.14). The first step in this process represents an empirically determined optimum needed to maximize the signal-to-noise ratio and quantify topological correspondences. The second reduces interimage variations due to lighting/exposure artefacts. The third allows the analysis to utilize spatially irregular 'regions of interest' as well as the more traditional rectilinear image boundaries.

Once *DAISY* had processed all images in a multispecies training set, a nonlinear discriminant space was calculated based on the *n*-tuple classifier (Lang, 2007; O'Neill, 2007). The proximate basis for this classification is a pair-wise comparison between RGB or monochromatic brightness values for each of the 32 × 32 pixel locations. The result allows each object in each training set to be placed into a multidimensional, distance-based ordination space whose character can be varied

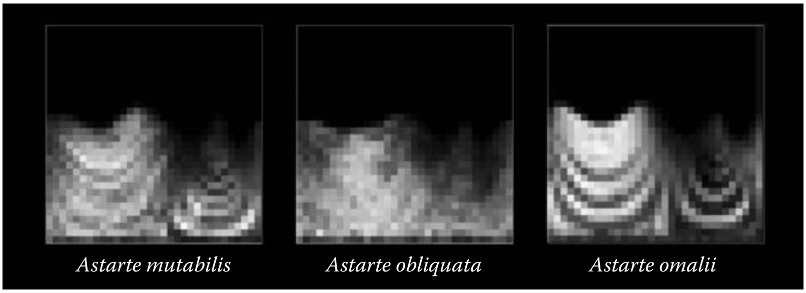


FIGURE 9.14 Thirty-two pixel polar coordinate image maps of the three *Astarte* specimens shown in Figure 9.8. The *DAISY* implementation of the PSOM-ANN uses these images to represent specimen morphology when operating in standard resolution mode. While these images may not appear to contain a great deal of information, *DAISY*'s record of performance in automatically identifying species from such images is remarkably good (see MacLeod et al., 2007a,b; ONeill, 2007, and references therein). For the raw image trial in this study, color images were used.

based on the estimated affinity (via normalized vector difference [NVD] correlation) between similarly processed images of unknown specimens and the training set array. It is this ability to modify the character of the base training set ordination that gives the *DAISY* implementation of the PSOM concept its adaptive quality (see earlier discussion). To achieve specimen identification, *DAISY* uses the computed affinity vector to project objects into the non-linear discriminant space.

The position of both known (training set) and unknown specimens in this space can be compared to the locations of training set objects using simple distance metrics. Cross-validation identifications were achieved by a multilayer strategy based on this discriminant space. The primary *DAISY* discriminator converts metric distances to a more robust rank measure, and assesses the eight nearest training set neighbours to the unknown object. Under this 'coordination' metric, the strength of group affiliation is measured by the number of neighbours belonging to the same group (e.g. all eight nearest neighbours in same group represent a coordination value of 1.00; six out of the eight neighbours represent a value of 0.75). In the current *DAISY* implementation, a coordination value of three or more is regarded as sufficient to associate the specimen with a group to a high confidence level.

If a specimen image cannot not be identified by the coordination metric, *DAISY* concludes the unknown specimen is not embedded within a group cluster. At this point a 'SILL' metric is used to determine whether it occupied a position at the edge of a group cluster. If the SILL metric cannot place the image within a known group, a 'first past the post' (FPTP) vote metric is used to determine the probable identity of objects in regions of the space containing members of more than a single training-set group by assigning the unknown object to the group of its nearest neighbour.

RESULTS

Three-Dimensional Relative Warps Analysis

Procrustes (GLS) superposition of the three-dimensional landmark data yielded four compact point clouds with landmarks 2 and 4 exhibiting a markedly more elliptical distribution than landmarks 1 and 3 (whose distribution is subcircular; see Figure 9.15a). This discrepancy indicates that there is a relatively strong component of anterior–posterior compression/expansion within these data and that the distribution of specimens along this compression/expansion gradient is continuous.

Singular value decomposition of these Procrustes shape coordinates produced six axes with non-zero singular values. Of these, three axes were required to represent 95 per cent of the observed shape variance within the pooled sample. When the original landmark configurations were projected into this three-dimensional linear shape space it was evident that the point clouds representing all three *Astarte* species exhibit broad overlap (Figure 9.15b).

Many practitioners discuss such plots as though they provide evidence that the shapes of the taxa in question intergrade morphologically. This is an erroneous interpretation of basic morphometric results. What this plot and subsequent plots of this type actually show is the degree to which between-groups separation can be achieved by sampling the morphology in the manner used to generate the raw

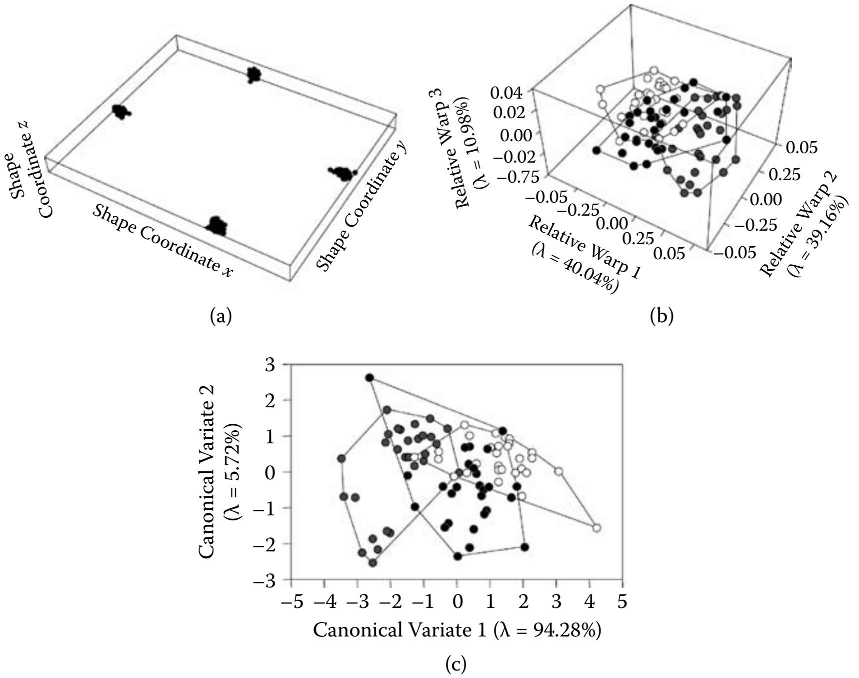


FIGURE 9.15 Results of three-dimensional landmark-based relative warps analysis. (a) Procrustes (GLS) superimposition of four landmarks collected from the pooled data-set of 90 *Astarte* specimens. (b) Perspective scatterplot of color-coded specimen positions in the shape space formed by the first three relative warp axes: *A. mutabilis* (black symbols), *A. obliquata* (grey symbols) and *A. omalii* (white symbols). This shape space represents 90 per cent of the total observed shape variation. (c) Scatterplot of color-coded specimen positions projected onto the two between-species canonical variate axes formed from analysis of scores on the first four relative warp axes (representing a total shape variance > 95%). Symbols as before. See text for description and interpretation.

data and portraying that separation within the first few axes of a high-dimensional space. This is not a subtle point. Nevertheless the absolute dependency of results on the data acquired and the axes selected to visualize the similarity relations is rarely discussed.

Although the interpretation of Figure 9.15b in terms of the ability of these four landmarks to distinguish among the three *Astarte* species appears clear, it is possible to use these data to produce a more complete picture of group distinctiveness: one that projects these morphologies into a space that is optimized with respect to the ratio of within-groups (*W*) to between-groups (*B*) variance. Accordingly, scores on the first three singular (= relative warp) axes were used as the basis for a CVA.

Because there are three species in the example data-set, only two discriminant axes are determined. Of these the first was markedly dominant, accounting for 94 per cent of the observed variation exhibited by the $W^{-1}B$ matrix. Projection of the original shape coordinates onto these discriminant axes (Figure 9.15c) confirmed the high level of overlap between the species-specific point clouds (compare

TABLE 9.1
Statistical Results for the Four-Point Three-Dimensional Relative Warp Analysis of *Astarte* Species Data

Wilks' Lambda Test (Rao's approximaton)		Pillai's Trace Test		Hotelling-Lawley Trace Test	
λ	0.355	Trace	0.698	Trace	1.666
$F_{\text{(Observed value)}}$	29.161	$F_{\text{(Observed value)}}$	23.325	$F_{\text{(Observed value)}}$	35.694
$F_{\text{(Critical value)}}$	2.424	$F_{\text{(Critical value)}}$	2.424	$F_{\text{(Critical value)}}$	2.461
DF1	4	DF1	4	DF1	4
DF2	172	DF2	174	DF2	102
ρ-value	<0.0001	ρ-value	<0.001	ρ-value	<0.001
α	0.05	α	0.05	α	0.05

Confusion Matrix

	A. mutabilis	***A. obliquata***	***A. omalii***	**Total**	**% Correct**
A mutabilis	18	4	8	30	60.00
A. obliquata	2	28	0	30	93.33
A. omalii	3	3	24	30	80.00
Total	23	35	32	90	77.78

to Figure 9.15b). However, in terms of recognizing the distinctiveness of these species, a four-landmark sampling strategy performed better than suggested by a cursory glance at the relative warps and CVA ordinations.

Calculation of the Wilks's λ, Pillai's trace and Hotelling–Lawley trace tests for equality of the group means relative to their dispersion all returned highly significant results (Table 9.1). This indicates that, despite the evident overlap of their shape distributions, these landmark data did pick out consistent differences between groups. In addition, the a posteriori assignment of specimens to group means using Mahalanobis distances suggests that 78 per cent of specimens could be placed with their correct group on the basis of this ordination. This test identified *A. mutabilis* as the most inconsistently identified species (= highest shape variability) when its shape was quantified by these landmark data (40% of specimens assigned to other species) and *A. obliquata* the most consistently identified (>7.00% of specimens assigned to other species). Overall, cross-tabulation results show that these simple landmark data contain sufficient information to assign slightly over three-quarters of individuals to the correct species.[10] Additional tests using truly unknown specimens would need to be undertaken to determine whether these discriminant functions could be used as generalized *Astarte* species classifiers. Regardless, these results are certainly sufficient to summarize the patterns present in this sample when those patterns are quantified using this set of three-dimensional landmark data.

Three-Dimensional Eigenshape Analysis

By using semilandmarks equally spaced around the *Astarte* valve's outline and tying the arrangement of these points to topologically corresponding landmark locations, much more information directly relevant to the characterization of these species can be collected. For this trial two landmark and 16 semilandmark points were located in three dimensions around the periphery of the valves with points 1 and 10 representing the landmark locations (Figure 9.10). The accuracy of this sampling procedure in representing the observed shape of the outline was greater than or equal to 95 per cent for each specimen in the sample (see MacLeod, 1999, for a detailed discussion of the sampling protocol used).

Because the outline was broken into two segments at the landmark points, different interpoint distances were used to characterize it. Results of the eigenshape analysis identified the posterior portion of the *Astarte* valve as the more complex from a shape analysis point of view, due no doubt to the relatively tight curve that characterizes the posterior portion of the valve's umbo due to the small number of landmarks that can be reliably located on the valve. As noted previously, none of these semilandmark points was allowed to shift or move during the analysis because to do so would have introduced error into the characterization of each specimen's outline and (by definition) no biological criteria are available for specifying the disposition of biologically corresponding structures along the valve margin (see Discussion section).

As with the three-dimensional landmark data, Procrustes (GLS) superposition of these three-dimensional outline data revealed interesting information about the sample. Figure 9.16 shows scatterplots of the superposition result in two and three dimensions. Comparing Figures 9.13 and 9.16 it is evident that the scarcity of geometric information from regions of the shape other than those represented by the landmarks led to a somewhat erroneous result. In Figure 9.16a it can be seen that, with the exception of the two landmark points, the distributions of all others exhibit substantial amounts of ellipticity. Rather than being transversely directed across the

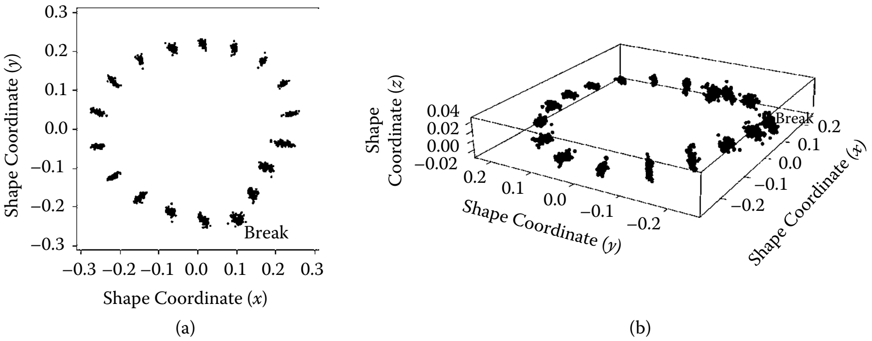


FIGURE 9.16 Procrustes (GLS) superimposition results for the 18 three-dimensional outline landmarks collected from the pooled data-set of 90 *Astarte* specimens. (a) Scatterplot of landmark point clouds in the x-y shape coordinate plane. Note overall tight clustering of point clouds and change in their geometry in the region of the beak-umbo. (b) Perspective scatterplot of the same data, this time showing the point-cloud distributions in three-dimensional space. Note important component of shape variation picked up by the morphometric z shape coordinate axis. See text for description and interpretation.

anterior–posterior axis, these semilandmark distributions were radially arranged in plan view (Figure 9.16a) and show characteristically different amounts of dispersion with marked degrees of incline to the commissural plane (Figure 9.16b).

Landmark 1 and the umbonal semilandmarks (2–5, 15–18) exhibited much greater amounts of scatter than the other boundary points. In addition, the major axes of their radially arranged point clouds were inclined at a high angle to the plane of the *x,y* shape-coordinate axes. This inclination was greatest in the commissural region, but varied along the value margin until the point clouds were, for the most part, parallel to the plane of the *x,y* shape coordinate axes in the valve's ventral region. These results suggest the greatest between-species differences in the outline lie in the valves' umbonal region.

Because of the higher geometric information content of the semilandmark-based sampling scheme, variance was spread over a larger number of singular axes (48). In terms of realized information content though, only four of these axes were needed to account for over 95 per cent of the observed shape variation. Projection of the original outline coordinates onto the first three of these axes (representing 93% of the shape variance) yielded a scatterplot in which *A. obliquata* was clearly distinguished from the other two species (Figure 9.17a). However, in this plot the outline shape intergradation between *A. mutabilis* and *A. omalii* appeared complete.

This latter result was confirmed by a CVA of the scores on the first four singular (= three-dimensional eigenshape) axes (Figure 9.17b). Once again, the first canonical variate axis was clearly dominant with an eigenvalue that represented 86 per cent of the total observed shape variance. Projection of the outline point sets onto both discriminant axes showed *A. obliquata* to occupy a well-defined region of the shape space with the *A. mutabilis* and *A. omalii* point clouds lying adjacent to each other, but also forming quasi-distinct regions. The point clouds of all groups, however, contained a small number of outliers that blur the otherwise clearly drawn morphological distinction between these species. Indeed, these results indicate three of the *A. mutabilis* and one *A. omalii* specimens appear to be misidentified.

Wilks's λ, Pillai's trace and Hotelling–Lawley trace tests all rejected the null hypothesis of no difference between group means ($p < 0.0001$). A Mahalanobis distance-based cross-tabulation result also indicated that 95 per cent of the training set specimens were able to be assigned to the correct species group (Table 9.2). Once again, *A. mutabilis* was the species with the most inconsistently distributed shape (10% misidentified specimens) and *A. omalii* the most consistent outline shape distribution (no misidentified specimens).

Eigensurface Analysis

A 10-point eigensurface sampling grid was used to quantify surface shape for the 90 valves comprising the example data-set. This resulted in specification of the two measured landmarks, and interpolation of 110 topologically corresponding semilandmarks for the pooled sample. These semilandmarks were arranged in a chevron-like pattern (Figure 9.18). The spatial density realized by this sampling scheme varied with the size of the specimen, but was on the order of a few millimeters.

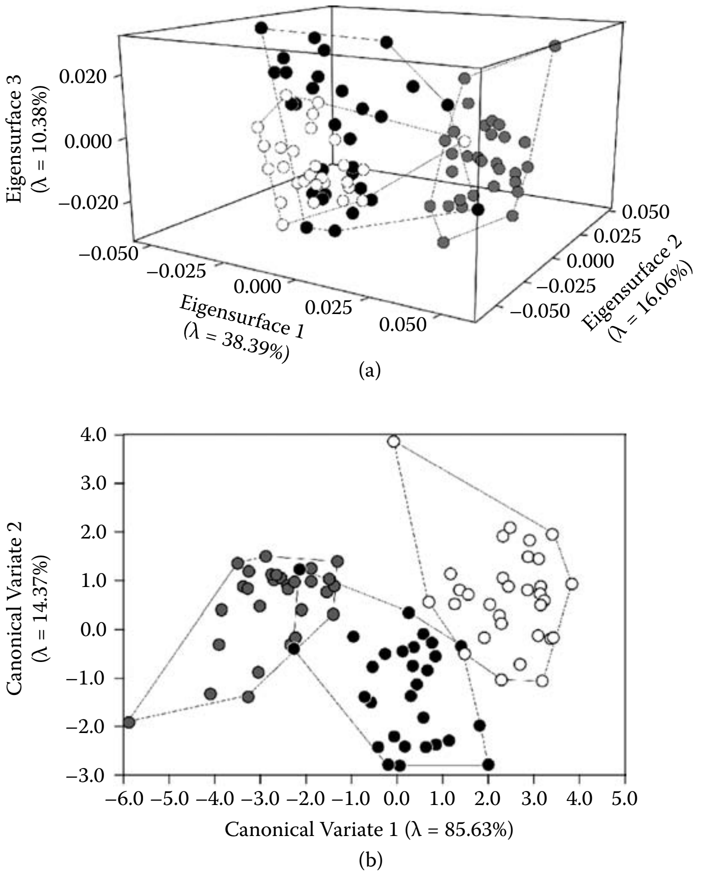


FIGURE 9.17 Results of three-dimensional outline-based eigenshape analysis. (a) Perspective scatterplot of color-coded specimen positions in the shape space formed by the first three three-dimensional eigenshape axes (symbols as in Figure 9.15). This shape space represents 65 per cent of the total shape variance. (b) Scatterplot of color-coded specimen positions projected onto the two between-species canonical variate axes formed from analysis of scores on the first four eigensurface axes (representing a total shape variance > 95%). See text for description and interpretation.

Figures 9.18a and 9.18b show results of a Procrustes (GLS) superposition of the surface grid points in two and three dimensions, respectively. A number of features characteristic of the pooled sample can be understood from inspection of these plots. Comparison of these figures shows that variation along the backbone semilandmark set is minimized in the lateral (*x*,*y*) direction, but extended along the transverse (right–left or *z*-axis) direction. This indicates the major component of shape variation in this region of the shells involves variations in the vault of valve surface.

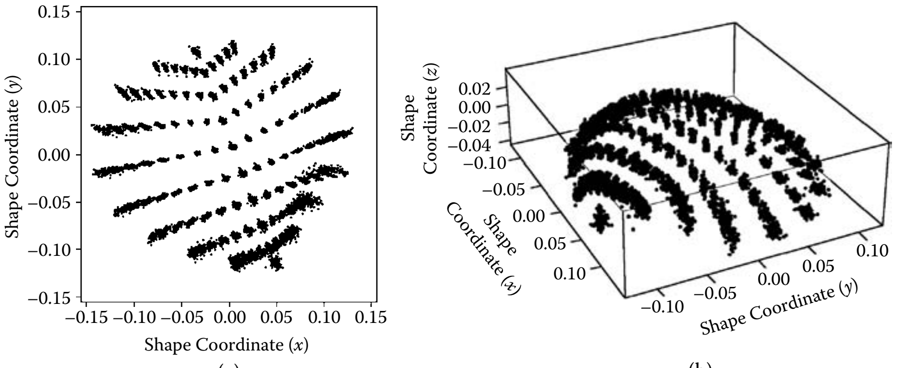
Moving away from this central region, the semilandmark point clouds were more diffuse and adopted a more laterally directed major axis. The more ventral regions

TABLE 9.2
Statistical Results for the 18-Point Three-Dimensional Eigenshape Analysis of *Astarte* Species Data

Wilks' Lambda Test (Rao's approximaton)		Pillai's Trace Test		Hotelling-Lawley Trace Test	
λ	0.102	Trace	1.254	Trace	5.326
$F_{\text{(Observed value)}}$	44.793	$F_{\text{(Observed value)}}$	35.692	$F_{\text{(Observed value)}}$	55.535
$F_{\text{(Critical value)}}$	1.994	$F_{\text{(Critical value)}}$	1.993	$F_{\text{(Critical value)}}$	2.018
DF1	8	DF1	8	DF1	8
DF2	168	DF2	118	DF2	118
ρ-value	<0.0001	ρ-value	<0.0001	ρ-value	<0.0001
α	0.05	α	0.05	α	0.05

Confusion Matrix

	A. mutabilis	*A. obliquata*	*A. omalii*	Total	% Correct
A mutabilis	27	2	1	30	90.00
A. obliquata	0	30	0	30	100.00
A. omalii	1	0	29	30	96.67
Total	28	32	30	90	95.56



(a) (b)

FIGURE 9.18 Procrustes (GLS) superimposition results for the 10-point eigensurface grid (= 112 three-dimensional landmarks) collected from the pooled data-set of 90 *Astarte* specimens. (a) Scatterplot of landmark point clouds in the *x*-*y* shape coordinate plane. Note the ease with which regions of high shape variability across the sample are picked out by the analysis, especially by the beak-umbo (= substantial variations in the tightness of the coil) and in the posterior ventral region. (b) Perspective scatterplot of the same data, this time showing the point-cloud distributions in three-dimensional space. Note important component of shape variation picked up by the morphometric *z* shape coordinate axis. In particular, note how much shape variability is expressed along the *z*-axis in each point cloud. This pattern of shape variation primarily reflects differences in the relative vaulting of the shell and would be invisible to a two-dimensional analysis. See text for description and interpretation.

of the shell exhibited laterally directed variation along an axis that diverged from the backbone axis, which itself approximates the mean coiling direction vector for each valve. This indicates the lateral regions of the valve exhibit more marginally directed shape variation with a decided preference for ventrally directed variation diverging from the midline in the ventral regions of the lateral margin.

This general pattern was disrupted, though, in the umbonal region, where the semilandmark point clouds exhibit very large relative dispersions. The localized, dorsally focused dispersion pattern for this umbonal region was such that boundaries between individual semilandmark point clouds were obscured, resulting in their mergence into a band of indistinctly structured shape variation when viewed over the pooled sample. One observation that can be made, however, is that shape variation in this region of the valve was directed transversely to a greater extent than laterally. Over the pooled sample, then, the umbonal region was identified as a locally important centre of shape variation in which both the distance from the coiling axis (= vault), tightness of the coil, and overall shape of the generating curve are all exhibiting copious variation. As for the structure of this variation, all point clouds appeared continuous in these plots, though this high shape variance may well obscure more complex patterns of species-specific shape variation.

Singular value decomposition of the Procrustes shape covariance matrix identified 54 axes with positive eigenvalues. Of these, 16 were required to summarize 95 per cent of the observed shape variance. A scatterplot of the first three singular vectors (= eigensurfaces), which together represent 93 per cent of the shape variation, showed all three species groups to be represented by relatively well-defined point clouds within this low dimensional shape space (Figure 9.19a).

Using projected scores of the original data onto each of the 16 independent surface shape variables as input into a CVA further refined the representation of species-specific shape differences across the sample. A plot of the resulting discriminant axes (Figure 9.19b) exhibited very well-defined species groupings with clear and large interspecies morphological gaps. This result corresponds well to the taxonomist's intuitive sense that these species exhibit clear morphological distinctions that can be easily recognized in terms of external shell morphology. The only exception was a single *A. mutabilis* specimen that projected to a position close to the margins of the *A. obliquata* point cloud. This outlier is the smallest sized specimen in the *A. mutabilis* group (specimen 5 of Plate 9.1), and as such does not exhibit many of the gross geometric attributes of larger sized *A. mutabilis* specimens (e.g. distinct vault, relatively narrow and tightly coiled umbo). An *A. omalii* specimen also plotted low and somewhat to the right of the main *A. omalii* point cloud. This specimen was not regarded as an outlier, but interpreted as a specimen that exhibits an atypically convex valve surface.

In terms of hypothesis tests, the Wilks's λ, Pillai's trace and Hotelling–Lawley trace statistics all rejected the null hypothesis of equality of species groups means (p < 0.0001; Table 9.3) but, as can be appreciated by comparing Figure 9.19 and Figures 9.15 and 9.16, did so by a decidedly greater margin. A cross-tabulation test for the training set data identified only one specimen–the smallest *A. mutabilis*–as being placed with an incorrect species group, yielding an overall accuracy ratio of 99 per cent.

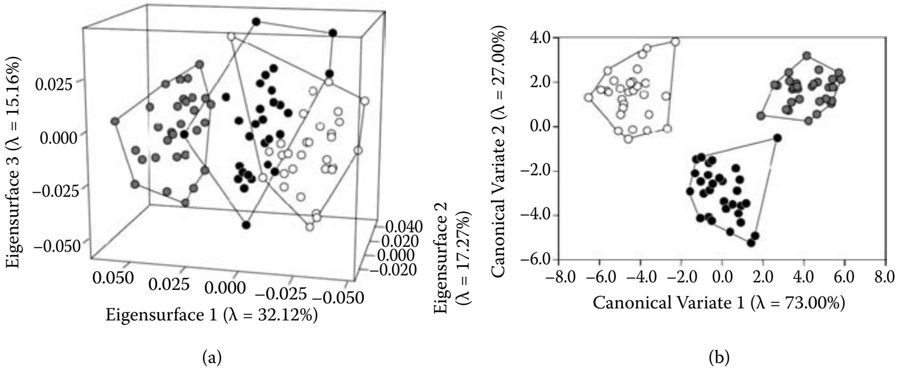


FIGURE 9.19 Results of eigensurface analysis (10-point grid). (a) Perspective scatterplot of color-coded specimen positions in the shape space formed by the first three eigensurface axes (symbols as in Figure 9.15). This shape space represents 65 per cent of the total shape variance. (b) Scatterplot of color-coded specimen positions projected onto the two between-species canonical variate axes formed from analysis of scores on the first 16 eigenshape axes (representing a total shape variance > 95%). Note much improved between-species separations. See text for description and interpretation.

With respect to the handling of aberrant specimens it is interesting to note that if centroid size is included in the variable set, the smallest *A. mutabilis* specimen remains an outlier from its species group with closer morphological affinities *to A. obliquata* (Figure 9.20a, Table 9.4). However, if the grid sampling resolution is increased from 10 boundary outline semilandmark points per outline segment to

TABLE 9.3
Statistical Results for the 10-Point Grid (112 Points) Eigensurface Analysis of *Astarte* Species Data

Wilks' Lambda Test (Rao's approximaton)		Pillai's Trace Test		Hotelling-Lawley Trace Test	
λ	0.013	Trace	1.786	Trace	19.326
$F_{\text{(Observed value)}}$	55.317	$F_{\text{(Observed value)}}$	38.171	$F_{\text{(Observed value)}}$	42.966
$F_{\text{(Critical value)}}$	1.612	$F_{\text{(Critical value)}}$	1.523	$F_{\text{(Critical value)}}$	1.537
DF1	22	DF1	32	DF1	32
DF2	154	DF2	146	DF2	124
ρ-value	<0.0001	ρ-value	<0.0001	ρ-value	<0.0001
α	0.05	α	0.05	α	0.05

Confusion Matrix

	A. mutabilis	***A. obliquata***	***A. omalii***	**Total**	**% Correct**
A mutabilis	29	1	0	30	96.67
A. obliquata	0	30	0	30	100.00
A. omalii	0	0	30	30	100.00
Total	29	31	30	90	95.56

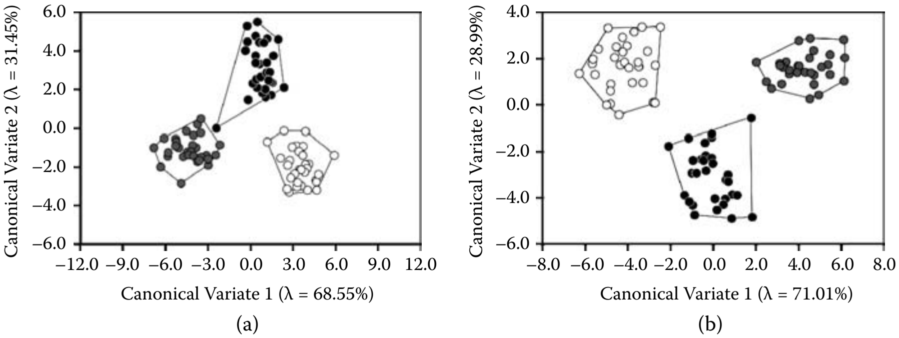


FIGURE 9.20 Alternative eigensurface results. (a) Scatterplot of color-coded specimen positions projected onto the two between-species canonical variate axes formed from analysis of scores on the first 16 10-point grid eigensurface axes (representing a total shape variance > 95%) and adding a 17th variable representing the (centroid) size of each specimen. Note overall tighter clustering of the *Astarte* species point clouds, with the exception of an anomalously small sized *A. mutabilis* individual. (b) Scatterplot of color-coded specimen positions projected onto the two between-species canonical variate axes formed from analysis of scores on the first 20 30-point grid eigensurface axes (representing a total shape variance > 95%). Note overall looser clustering of the *Astarte* species point clouds, with the 'pulling back' of the anomalously small sized *A. mutabilis* individual to the *A. mutabilis* point cloud. This may suggest the characteristic *A. mutabilis* features in small sized individuals exist at very fine spatial scales. Note also arbitrary direction of canonical variate axes–a well-known feature of eigenanalysis. See text for description and interpretation.

TABLE 9.4

Statistical Results for the 10-Point Grid (112 Points) Eigensurface and Size Analysis of *Astarte* Species Data

Wilks' Lambda Test (Rao's approximaton)		Pillai's Trace Test		Hotelling-Lawley Trace Test	
λ	0.014	Trace	1.753	Trace	16.205
$F_{\text{(Observed value)}}$	66.636	$F_{\text{(Observed value)}}$	63.193	$F_{\text{(Observed value)}}$	70.421
$F_{\text{(Critical value)}}$	1.669	$F_{\text{(Critical value)}}$	1.669	$F_{\text{(Critical value)}}$	1.685
DF1	18	DF1	18	DF1	18
DF2	158	DF2	160	DF2	128
ρ-value	<0.0001	ρ-value	<0.001	ρ-value	<0.0001
α	0.05	α	0.5	α	0.005

<u>Confusion Matrix</u>

	A. mutabilis	*A. obliquata*	*A. omalii*	Total	% Correct
A mutabilis	29	1	0	30	96.67
A. obliquata	0	30	0	30	100.00
A. omalii	0	0	30	30	100.00
Total	29	31	30	90	95.56

TABLE 9.5
Statistical Results for the 30-Point Grid (112 Points) Eigensurface and Size Analysis of *Astarte* Species Data

Wilks' Lambda Test (Rao's approximaton)		Pillai's Trace Test		Hotelling-Lawley Trace Test	
λ	0.012	Trace	1.760	Trace	17.430
$F_{\text{(Observed value)}}$	56.029	$F_{\text{(Observed value)}}$	52.046	$F_{\text{(Observed value)}}$	60.407
$F_{\text{(Critical value)}}$	1.612	$F_{\text{(Critical value)}}$	1.611	$F_{\text{(Critical value)}}$	1.626
DF1	22	DF1	22	DF1	22
DF2	154	DF2	156	DF2	128
ρ-value	<0.0001	ρ-value	<0.0001	ρ-value	<0.0001
α		α		α	

Confusion Matrix

	A. mutabilis	*A. obliquata*	*A. omalii*	Total	% Correct
A mutabilis	30	0	0	30	100.00
A. obliquata	0	30	0	30	100.00
A. omalii	0	0	30	30	100.00
Total	30	30	30	90	100.00

30 points per segment (specifying the use of 856 total grid points to achieve a spatial resolution of well below 1.0 mm), subtle similarities between this specimen and other *A. mutabilis* specimens are picked up, causing the former to 'pull back' closer to the overall *A. mutabilis* point cloud, Indeed, use of this sampling-grid resolution causes all species-group point clouds to increase their variance. As a result, the formerly aberrant specimen migrated to a decidedly closer position–as measured by the Mahalanobis distance–relative to the *A. mutabilis* centroid than to the centroids of the other two species (Figure 9.20b, Table 9.5). This agrees with the qualitative inspection of this specimen, which does exhibit a distinctly 'mutabilis' character despite being different from all other *A. mutabilis* specimens in the sample. While acknowledging the anecdotal nature of this case it would seem that–at least in terms of eigensurface analysis–increasing the spatial resolution of the shape data did a better job in resolving this ambiguity than including different types of data.

PSOM-ANN Analysis

Three data-sets were submitted to the *DAISY* PSOM-ANN implementation, with training image sets representing (1) color (RGB) images of the actual shells, (2) greyscale images of the virtual shells represented by the adaptively meshed three-dimensional data used to construct the eigensurface grids (see earlier discussion and Plates 9.4–9.6) and (3) greyscale images of the 10-point grid data used for the eigensurface analysis (Plates 9.7–9.9). The first of these data-sets represents a 'warts and all' test of the taxonomic information contained in the raw images, at least insofar as this information can be captured by the 32 × 32 pixel sampling grid used by the

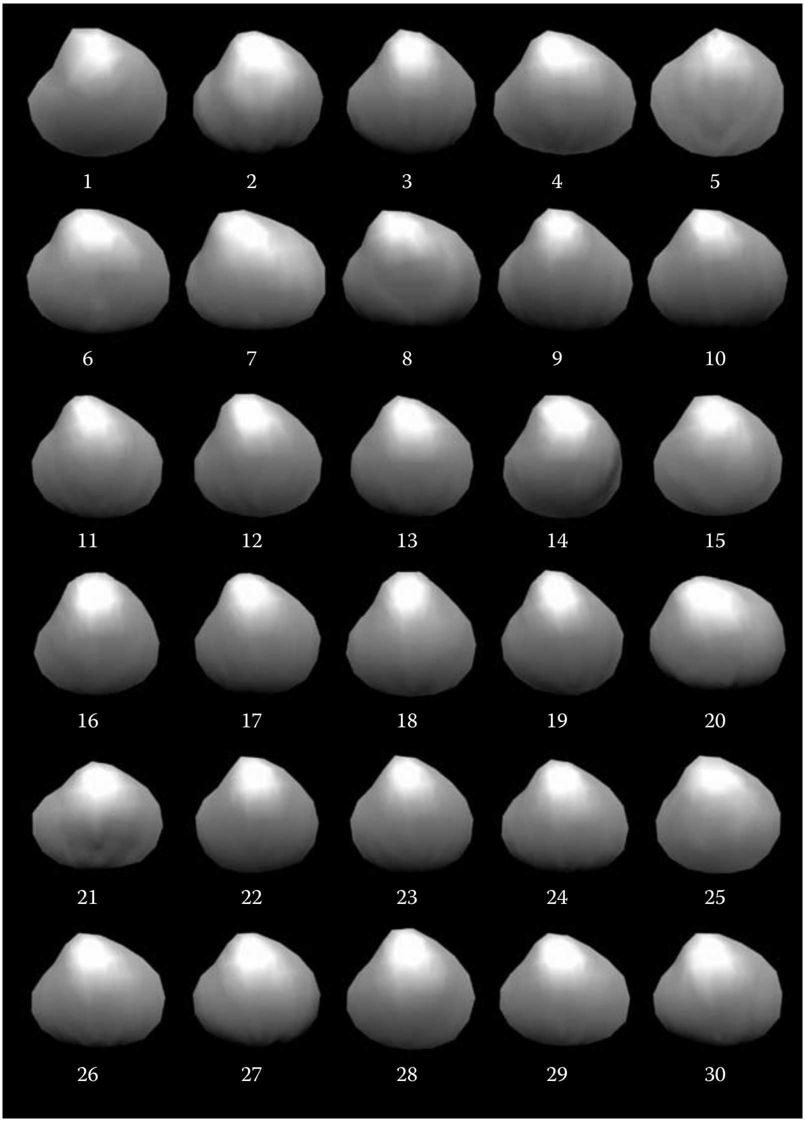


PLATE 9.7 Virtual models of *Astarte mutabilis* specimens created using the 10-point grid data that was employed in the eigensurface analysis. These images were also used as input into the *DAISY* PSOM-ANN analysis.

DAISY application for comparative analyses. This sampling scheme represents the shell morphologies as a data-set containing more information than any of the three comparable morphometric data-sets used in the analyses described before. However, these data were also more complex because these valve surfaces contain a number of artefactual defects (e.g. color variations, shadows, preservation artefacts, curation artefacts, damage). While not obscuring the taxonomic utility of the values for

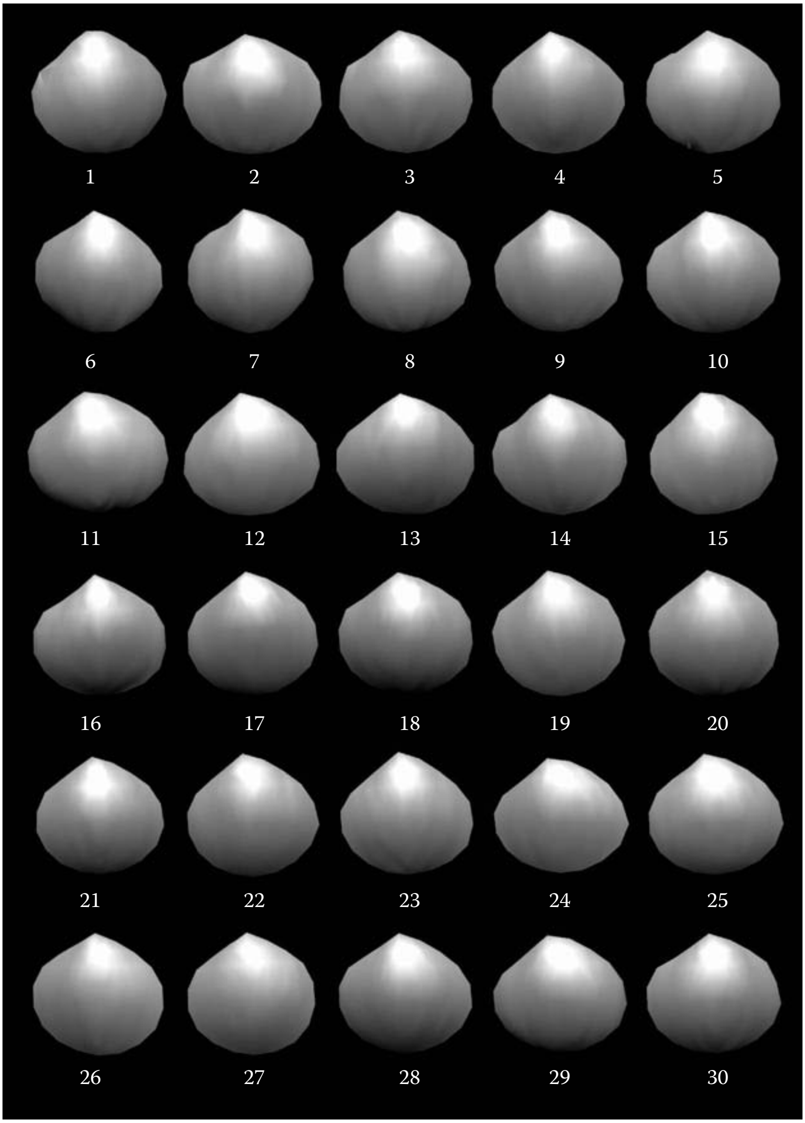


PLATE 9.8 Virtual models of *Astarte obliquata* specimens created using the 10-point grid data that was employed in the eigensurface analysis. These images were also used as input into the *DAISY* PSOM-ANN analysis.

qualitative evaluation by a trained taxonomist, such variations inevitably complicate the problem of group characterization for automated systems.

The second data-set remedied aspects of the extraneous morphological information that characterized the raw images by standardizing color and lighting artefacts as well as be eliminating small spatial scale differences. This data-set represents an estimate of results that might be achieved by relatively simple processing of the

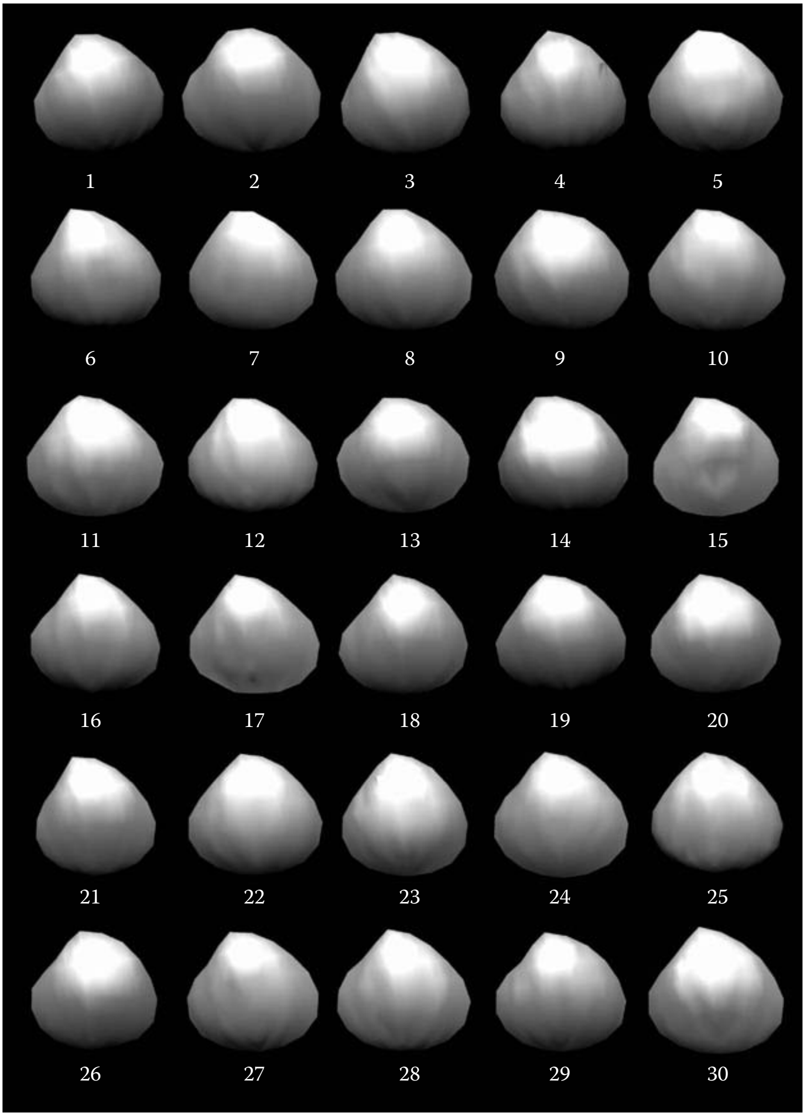


PLATE 9.9 Virtual models of *Astarte omalii* specimens created using the 10-point grid data that was employed in the eigensurface analysis. These images were also used as input into the *DAISY* PSOM-ANN analysis.

valves via three-dimensional scanning. The third data-set reduced the spatial resolution of the three-dimensional scanned data dramatically, but preserved the overall shape of each valve in a manner that the eigensurface results had previously identified as being able to represent the greatest between-groups shape differences.

It is important to remember that the data submitted to the *DAISY* system are not true three-dimensional data (as were used in each of the morphometric data analyses). Rather, *DAISY* processes images that represent the third data dimension

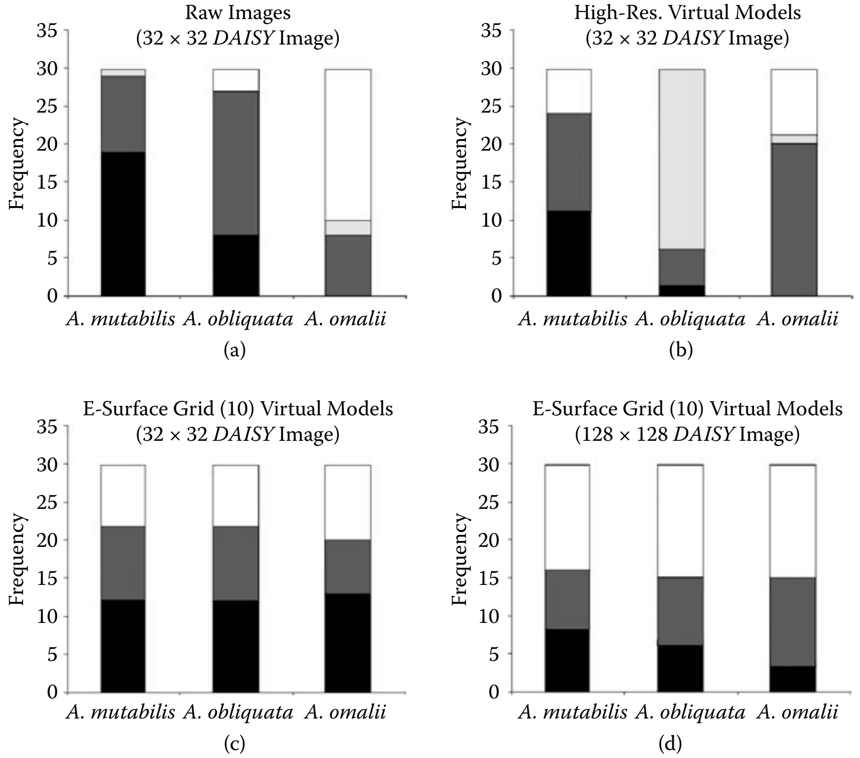


FIGURE 9.21 Cross-tabulation results of *DAISY* analysis. In all plots, white: frequency of correct training set identifications passed under the coordination test; light grey: frequency of correct training set identifications passed under the SLL test; dark grey: frequency of correct training set identifications passed under the FPTP test; black: frequency of incorrect training set identifications. See text for description of data-sets and interpretations of results.

as patterns of relative lightness and darkness–topologically high and low regions, respectively–that interact with color, texture, compositional and lighting variation in complex ways. This having been said, raw images of specimens are routinely used to illustrate taxonomic concepts and document taxonomic identifications throughout the systematic literature. The two virtual image data-sets variously normalize the specimen images for complications in the portrayal of this third dimension in the raw image set, thus boosting the amount of true topological information presented for analysis.

Results of cross-tabulation analyses of the *DAISY* training sets are shown in Figure 9.21. These frequency diagrams not only tabulate the number of correct/incorrect *post hoc* identifications, but also provide an indication of the quality of or confidence in identifications referenced to the coordination-SLL-FPTP scale.

For the raw images (Figure 9.21a) 63 (70%) of the training-set images were placed in the correct groups. As with the morphometric results, *A. mutabilis* was identified as the most variable group (37% correct identifications) and *A. omalli* as the least variable (100% correct identifications). In terms of confidence in group assignments, both *A. mutabilis* and *A. obliquata* valves were characterized by relatively high levels of dispersion within the image-based feature space. In no case was the image of any *A. mutabilis* specimen placed in the discriminant space such that its three nearest neighbours were also *A. mutabilis* specimens. Similarly, only three *A. obliquata* specimens were embedded in successive jackknifed *A. obliquata* distributions such that the cut-off coordination limit of 3 was satisfied. In both these species the majority of the correct species group assignments were made using the FPTP criterion of assigning the image to the group represented by the single nearest neighbour (coordination value = 1).

These results stand in contrast to those obtained for *A. omalii*, 67 per cent of whose identifications passed the much more stringent coordination test. Indeed, the mean coordination value for *A. omalii* specimens (6.00) was also noticeably higher that the mean coordination value for *A. obliquata* (3.67). These results are consistent with inspection of Plates 9.1–9.3, which show both the *A. mutabilis* and *A. obliquata* training sets to be much more inconsistent in terms of color variability and the presence of various artefacts than *A. omalii.*

Normalization of the raw images for these sources of variability using meshed representation of high-density three-dimensional scanned point clouds (Figure 9.21b) resulted in an overall improvement in the number and quality of correct identifications. By using images of the virtual objects as input to *DAISY* the overall number of correct identifications of training set images rose to almost 80 per cent. The *A. mutabilis* results underwent a particularly dramatic improvement with overall correct identifications rising to 66 per cent, six of which were of the highest quality (mean coordination value: 4). Even the lower quality FPTP results improved for this group, rising from a previous mean value of 0.60–0.94.

The *A. obliquata* results underwent even greater levels of improvement, especially in terms of the overall identification confidence. Using the virtual images, an additional seven specimens were able to be assigned to the *A. obliquata* group using the cross-tabulation test. Of these 29 specimens, 24 met the relatively stringent SLL identification criterion. However, cross-tabulated identification results for *A. omalii* were markedly lower for the virtual image data-set compared to the raw images. The number of correctly assigned *post hoc* identifications dropped from 30 to 22. Of these, over half were made using less stringent quality FPTP criterion (mean FPTP value: 0.86). Nevertheless, for these three data as a whole, use of the three-dimensional virtual images resulted in almost doubling the number of high-quality, correct cross-tabulated identifications of training set images.

Use of virtual images constructed from the 10-point sampling grid that produced the most distinct group-level distributions in the morphometric analysis provided perhaps the most interesting results of the investigation. The purpose of undertaking this test was not only to facilitate direct comparison with the eigensurface-based morphometric results, but also to examine whether use of three-dimensional scanning technology could improve the performance of ANNs. The virtual valves shown in Plates 9.7–9.9 represent images of shapes that are known to embody the geometric distinctions between these species and that also have been normalized for extraneous sources of variation due to fine-scale textural variations, color variations and lighting variations.

Nonetheless, the *DAISY* PSOM-ANN implementation was not able to parse these images into characteristic shape groups (Figure 9.21c). Using a 32 × 32 sampling grid, only 32 (58.89%) of these training set images were placed into the correct species groups in the cross-tabulation test. Of these, only 26 (28.89%) images were identified with high levels of confidence by the coordination test, though the mean coordination value (5.26) was relatively high. Mean FPTP values (0.98) were also high. Still, the inescapable conclusion is that standardizing the image of the object was, by itself, insufficient to improve identification accuracy. Indeed, the number of high-quality identifications for these data-sets of highly processed images is identical to the number obtained for the raw images and much less than the number obtained using the high-resolution virtual images (compare Figures 9.21a–9.21c).

There are two strategies by which the performance of ANNs can be improved without raising the resolution of the virtual images or changing the network's fundamental design. The first is to increase the size of the training sets (see Bishop, 1996, and Lang, 2007, for theoretical discussions and MacLeod et al., 2007b, for a practical example). Since the samples used in this investigation do not allow for increasing species sample sizes significantly and since this is a well-known feature of ANNs, a direct demonstration of this effect was foregone.

The second strategy for improving performance involves raising the information content of image itself, by increasing the resolution of the image. This operationally differs from increasing the resolution of the data used to construct the virtual specimen, because it involves an increase in the size of the ANN used to assess and characterize morphological variation. To explore the effect of this strategy a series of tests using grid sizes of 64 × 64, 96 × 96 and 128 × 128 were conducted. Only the last of these is reported here because was the highest resolution analysis of any (morphometric or ANN) attempted and returned the best results for any ANN trial.

Note that whereas the 10-point and 30-point morphometric grids consisted of data-sets of 345 and 2577 points, respectively, the 128 × 128 *DAISY* grid based its analysis on a data-set that consisted of 16,348 discrete locations over the image frame. But despite this higher resolution, performance in the cross-tabulation test was decidedly *less* impressive than for either the eigenshape-CVA or the eigensurface-CVA results. In total 73 (81.11%) of the images presented in Plates 9.7–9.9 were placed in their correct species groups by the 128 × 128 *DAISY* analysis. Forty-four (48.89%) were identified to a high confidence level (all coordination, mean value: 5.26). This is a large improvement over the 32 × 32 results and would certainly render an ANN-based approach useful for routine taxonomic identifications (see MacLeod et al., 2007a,b). Controlling the images of these bivalve shells closely for texture, color and lighting variation resulted in only marginal improvements in the ability of the PSOM ANN to characterize and correctly identify species, whereas increasing the spatial resolution of the data used to construct the virtual image analysed by the PSOM ANN resulted in greater accuracy gains.

Based on these results it appears both strategies can, in principle, improve the ability of ANNs to characterize and identify biological species. Both strategies are also easy to implement using contemporary three-dimensional scanning technology. Neither of these strategies, though, matched the performance of the eigensurface-CVA trials, no doubt because of the superior geometric information content of the high geometric information content of the data that served as the basis for that analysis. Of course, an alternative automated specimen identification strategy would be to submit the three-dimensional eigensurface grid-coordinate data directly to ANN processing. However, in light of the results achieved using linear multivariate discrimination, this step was deemed unnecessary.

DISCUSSION

The results presented here illustrate the use of new technologies and data analysis concepts in morphological data analysis–especially in the context of species identification–as well as pointing up areas where important future research problems lie. They also both clarify and challenge a number of basic assumptions about the use of morphological data in both systematic and morphometric contexts.

First, the advantages of adopting a quantitative approach to morphological data analysis are well illustrated by the example analysis. After almost 50 years of publications on morphometric approaches and their application to a wide variety of systematic problems, it is discouraging to acknowledge that such methods remain the exception in most systematic studies. Adams et al. (2004) showed that the number of morphometric articles has increased dramatically over the 20 years since the Michigan Morphometrics Workshop where geometric morphometrics was first introduced. What Adams et al. neglect to point out is the existence of a sizeable body of literature explicitly arguing that morphometric methods have no place in contemporary systematic studies (see MacLeod, 2002a, for a review). This stance is illogical, for it implies that imprecise, irreproducible, qualitative observations of complex morphological variation patterns should always be preferred over precise, reproducible measurements and summaries of the same structures and patterns (see MacLeod et al., 2002; MacLeod, 2004, 2005; MacLeod et al., 2007b). Additionally, over-reliance on qualitative observations has led to endless and irresolvable disagreements over the definitions of morphological character states, assignments of character states to species, species concepts, species identifications, etc. This imprecision and lack of reproducibility not only compromises primary systematic data, but also affects (to varying degrees) all the secondary and tertiary systematic data analyses and results such data underpin. The morphometric investigation of morphological data by systematists can improve this situation now and will be able to do a better job in the future as three-dimensional data and data analysis procedures become more common.

Turning from general issues to the specifics of the example analyses, here we can glimpse both the promise of, and the problem with, morphometric results. Despite the unexpectedly good results for statistical and cross-tabulation tests obtained from the three-dimensional landmark and three-dimensional outline data-sets, the CVA ordination plots make it clear that both badly misrepresent the information present in the actual specimens. Even to the untutored taxonomic eye, Plates 9.1–9.3 show that all three species are quite distinct in terms of their characteristic morphology, with little apparent morphological intergradation between species morphologies.

However, the CVA ordination plots shown in Figures 9.15c and 9.17c suggest broad regions of overlap exist among all three species.

In retrospect it may seem obvious that the results of the landmark, outline and surface data analyses must be judged within the contexts of those portions of the overall morphological data-set they actually sample. Unfortunately, this is precisely what all too often fails to happen in practice. The separation between the *Astarte* species point clouds in Figures 9.15c and 9.17c was incomplete, not because the morphologies themselves or the species concepts taxonomists have erected for these taxa are indistinct, but because qualitative taxonomists routinely access information about the character of the valve surface morphology that neither a few landmark points nor a more densely spaced set of boundary outline semilandmark points can capture even when these points are sampled within a three-dimensional context. If the problem the systematist is interested in requires him or her to focus on the measurement of these landmarks or outlines, such plots are correct and fully indicative of biological reality. However, if the question is of a more general nature (e.g. whether the shapes of these species' valves differ) or pertains specifically to the character organismal surfaces, such as Figures 9.15c and 9.17b are incomplete at best, misleading at worst.

Of course, the grid-based ordinations presented in Figures 9.19b and 9.20b also show that surface-based approaches must be referenced to specific spatial resolutions. In this sense, 'shape' has an irreducibly scale-dependent or fractal character. Regardless, the availability of three-dimensional surface-based approaches emphasizes the need to match (and the practical advantages of matching) the data acquired and scale of data acquisition to the systematic/geometric problem at hand.

Reflecting on the strengths and weaknesses of landmark and outline data as illustrated in this investigation, it seems clear that landmark data *per se* should be used when the problems of interest involve the relative distribution of structures that can be defined by point locations or when the shape of the specimens and/or structures are so simple and well defined that they can be adequately characterized by a small number of landmarks. Outline data should be used when the problems of interest pertain to the shape of peripheries or curves too complex or ill-defined to be susceptible to landmark analysis, but the limits of which can be delineated by accurate landmark locations. More-or-less everything else should be analysed by appropriately resolved three-dimensional surface data in the form of topologically corresponding three-dimensional grids because these representations of shape provide the most complete assessment of the total morphological information available and–even more importantly–because they embody a quantified (and quantifiable) facsimile of the data accessed routinely by qualitative systematists–taxonomists. Anything less can result in the production of morphometric results that, to a greater or lesser extent, fail to adequately represent the geometries of the structures of interest.

The evident failure of either three-dimensional landmark or three-dimensional outline results to 'capture' the clear geometric distinctions that exist between these bivalve shapes also, I believe, explains why morphometrics has not made greater inroads into mainstream taxonomy. Taxonomists are conservative by training and by nature. Despite the elegance of the mathematics involved, if the ordination plots that represent morphological data do not exhibit between-group distinctions of the same order that taxonomists 'see' in front of them, they are bound to treat such results with suspicion, and rightly so. In this sense, the lack of a strong connection between systematics and morphometrics over the issue of morphological analysis may be nothing more than a failure of morphometricians to 'sense' their shapes in manners that are comparable to those of classic comparative morphology. Once this has been done, though, the amount of useful information that can be obtained by morphometric approaches far outstrips that which can be obtained by qualitative morphological analysis.

The morphometric results obtained from the example *Astarte* data also clarify a number of other important issues. In terms of morphometric technique, eigensurface analysis (and related procedures) now more-or-less complete the geometric morphometric toolset that began to be developed with the advent of shape coordinates (Zahn and Roskies, 1972; Siegel and Benson, 1982; Bookstein, 1986). Some work still needs to be done to allow eigensurface analysis to take advantage of the location of landmarks inside the object outline and to implement the analysis of matched grids from different sides of the same object (= a complete virtual object), but these are relatively simple matters. When completed, eigensurface analysis will be able, in principle, to support rigorous geometric analyses of virtual objects of any level of complexity while taking full advantage of the constraining information contained in biological landmarks.

Eigensurface analysis elegantly demonstrates the need to combine the information provided by biological landmarks with the topological information provided by semilandmarks to achieve complete quantitative representation of biological form suited to detailed quantitative analysis. In addition, eigensurface analysis serves to further illustrate the utility of semilandmarks as shape descriptors. The representation of surfaces–just like the representation of curves or curve segment–is inextricably bound to the points used to characterize the surface.

In addition to clarifying how landmark, outline and surface-based morphometric methods should be used in systematic–taxonomic contexts, the results produced by this study raise interesting questions about the currently popular method of 'sliding semilandmarks' (Sampson et al., 1996; Bookstein, 1997; Bookstein et al., 1999) as an alternative to equally spaced semilandmarks as outline descriptors. Under the current sliding landmark procedure, shapes represented by combinations of landmarks and semilandmarks are superimposed using data from Type 1 or Type 2 landmarks according to the standard Procrustes approach. A reference shape is then selected and the outline semilandmarks slid along tangents to the curve at the position of each semilandmark until either a minimum bending energy configuration has been achieved or the target-shape semilandmarks occupy positions aligned with chords perpendicular to tangents at each reference semilandmark. The purpose of this adjustment is to bring the semilandmarks into positions of maximum shape correspondence (= minimum bending energy) with the reference form.

Although the mathematics behind this procedure are interesting, the effects of its use are counterintuitive. Maximum correspondence between semilandmarks would appear to be a desirable property for any morphometric data analysis to aspire to. However, under the sliding-landmark procedure, topological correspondence can be achieved only by *changing the empirically measured shape,* thus violating the principle of maintaining strict topological correspondence between the mathematical representation of the specimen and the specimen's actual morphology. For some geometric morphometricians, such shape deformation is allowable because they regard semilandmarks as having a different ontological status from landmarks. However, this ignores the fact that, in many instances, the criteria for placing so-called landmarks on a form differ little from the criteria used to place semilandmarks. In other words, practically speaking, there is often little difference between landmarks and semilandmarks in terms of their placement rules. This is certainly the case for three of the four landmarks used in the *Astarte* example and this example is not different from many landmark-based analyses that have been published in the systematic and morphometric literature (e.g. see chapters in Marcus et al., 1993). If there is no practical difference between the definitions of a landmark and a semilandmark in a given study, it makes little sense to treat these geometric constructions differently during data analysis–especially when doing so changes the geometry of the empirical shapes under consideration.

For curves densely (= accurately) represented by semilandmarks, sliding will result in only minor and most likely inconsequential adjustments. For curves represented by sparse sets of semilandmarks, the problems inherent in the sliding-landmark approach are more serious. First, specification by sparse semilandmarks cannot adequately represent the complexity of most biologically interesting outline curves and/or curve segments. Second, the convention of adjusting sparse sets of sliding landmarks along tangents to outline curves inevitably results in substantial distortion of the curve's representation, thus divorcing the analytic result to a greater or lesser extent from the very specimens whose shape the analysis was undertaken to assess. Third, the convention of allowing outline descriptors to slide along tangents until a configuration of 'minimal bending energy' or 'maximum radial shape correspondence' is achieved is both artificial and arbitrary from the standpoint of most biological hypotheses.

The standard justification for this procedure appears to be that 'if you don't know where the semilandmarks are supposed to be located you should let the analysis estimate their position for you' (see Zelditch et al., 2004). From a systematic–taxonomic point of view this justification seems wholly unsupportable when 'letting the analysis decide for you' means the analysis 'decides' what shapes are being analysed and is not particularly concerned with the fact that the resultant shapes bear no precise or consistent relation to actual specimens. This will be especially problematic for two-dimensional curves, for many of these lack the critical third dimension of information necessary to help ensure even approximate topological correspondence. Interestingly, the sliding of semilandmarks to corresponding positions may be more appropriate for the three-dimensional case because of its inherently higher geometric information content. In such cases I suspect it will be more fruitful–and justifiable–to constrain sliding using the Sampson et al. (1996) 'perpendicular to reference shape tangents' shape minimization criterion, since it may not be possible to achieve a true minimization of bending energy configuration owing to the large number of possible semilandmark displacements–but only if the semilandmarks are slid along the surface itself rather than along tangents to that surface. The data analytic problems associated with sliding semilandmarks highlight the need to be mindful of both the geometric and biological dimensions of a morphometric problem and to remember that the purpose of morphometrics is to provide tools to help explain biology (not geometry).

It is also interesting to note the differences between CVA results for the 10-point and 30-point eigensurface grids (Figures 9.19b and 9.20b). Both ordinations are based on highly accurate representations of valve form. However, the lower resolution 10-point grid result is slightly better in terms of minimizing intraspecific variance and maximizing interspecific difference than the higher resolution 30-point grid results. To restate this in terms of a popular palaeontological research programme, the 10-point grid results exhibit a greater morphological disparity than the 30-point grid results, not to mention the three-dimensional landmark and three-dimensional outline results.

Metrics of morphological disparity have been used to study patterns of morphological diversification over time (see Foote, 1989, 1991, 1993, 1996). Nevertheless, geometric metrics of disparity are inevitably tied to data type (e.g. two-dimensional morphometric disparity measures will not necessarily be congruent with three-dimensional morphometric disparity measures; see MacLeod, 1999). The *Astarte* results indicate that geometric disparity measures exhibit a scale dependence that has not been taken into consideration to date in the summary of morphological diversification patterns over evolutionary time.

Semiquantitative analyses of disparity (e.g. quantitative analysis of character-state data; see Foote, 1996) have the ability to assess data at various spatial scales. However, the results of such analyses are 'scale free' and cannot be used to adequately summarize the morphological (= geometric) patterns of variation because the very concept of morphology is logically tied to the concept of scale, both in terms of raw size-shape analysis and size-free shape analysis. Eigensurface analysis represents an interesting alternative approach to disparity analysis, in the sense that it leads to a more thoughtful and rigorous documentation of morphology and produces summaries of morphological diversification patterns that emphasize the importance of scale dependence.

Finally, it is necessary to return to the core issue of this chapter and consider the implications of the results presented here with respect to fully automated specimen identification. Both the 10-point and 30-point eigensurface grids were able to support essentially perfect discriminant spaces for the training set data. Between-groups separations were such that, assuming these samples are representative of the Coralline Crag and Red Crag localities, there can be little doubt such spaces would provide a generally useful basis for identifying unknown specimens, at least for samples recovered from those localities. None of the two-dimensional image sets submitted to the *DAISY* PSOM-ANN implementation was able to achieve a comparable level of identification accuracy, though all morphometric and most ANN results were more consistent than average levels of identification accuracy achieved by the traditional qualitative approach to specimen identification (see MacLeod, 1998; Culverhouse, 2003, 2007).

As with the comparison of different morphometric approaches to morphological sampling, the *DAISY* results suffered from the fact that the image data do not accurately represent the three-dimensional nature of the valve surface. On the other hand, the collection of such data is time consuming, a factor that must be taken into consideration when judging the performance of automated specimen recognition systems. Here, the best rule of thumb is to carefully consider the needs of the problem (Is 100% accuracy essential?), the nature of the specimens (Are three-dimensional data needed to accurately represent the morphology?) and the availability of training set specimens. (ANNs work well but usually require much larger training sets than morphometric methods; see MacLeod et al., 2007b.)

For the example data-set, the most compelling comparison is that between *DAISY* results and the eigensurface results since the grid points roughly correspond to image 'pixel' values. Image pixels may be able to capture enough of the three-dimensional morphology to be of use for identification of some groups. These will likely be better than sparse landmark data and generally the same as outline data, depending on the quality of the input images and the size of the training set. Nevertheless, three-dimensional grid data will always be better for representing three-dimensional morphologies. Figures 9.15–9.18 show how much better they can be. Highly accurate automated systems designed to analyse digital images are available now (see MacLeod, 2007c). Automation of scanning systems to extracting and utilizing three-dimensional grid data will be more challenging to construct. But such systems can be constructed. Whether the costs associated with doing so are prohibitive enough to restrict the three-dimensional grid approach to specialist investigations remains to be seen. Addressing this issue will require setting up a specific interdisciplinary research project to further develop and combine the necessary technologies into a prototype system to be tested using a variety of real-world examples.

SUMMARY

There is an undeniable need to quantify systematic data and data analyses in order to improve their accuracy and reproducibility, and in order to enable them to be compared to other types of data–most of which have long been expressed in quantified form. The lack of an ability for systematists to report their data in quantitative terms is at least partly responsible for the lack of support systematics receives from many allied disciplines. To the extent that systematics has been, is now and will continue to be primarily concerned with comparative morphological investigations–or at least investigations in which the ability of systematists to identify organisms based on morphological criteria plays an important role–morphometric approaches will prove increasingly useful. Although the scope of both systematics and morphometrics is large, with plenty of opportunity for various types of data and approaches, perhaps the most important single contribution contemporary morphometrics can make to contemporary systematics is through undertaking research designed to produce automated computer vision/intelligence systems that can reliably identify species based on morphological criteria.

Previous studies (MacLeod et al., 2007a) have shown that both landmark-based morphometric analysis and ANN-based image analysis strategies can produce accurate identification results based on the assessment of specimen form in two dimensions. The example analyses presented in this article demonstrate that, as good as two-dimensional summaries of morphological variation can be, precise quantification of taxonomic species concepts that have been developed through qualitative inspection of three-dimensional objects will likely require three-dimensional data sampling and analysis strategies to realize fully.

In order to facilitate comparisons with traditional morphometric procedures (landmark sets, outlines) and with the manner in which systematists have traditionally undertaken comparative morphological analyses, a new method of shape analysis–eigensurface analysis–has been developed. Eigensurface analysis extends the eigenshape approach to the quantification of biologically interesting outlines and outline segments to three-dimensional surfaces. Comparisons of the ability of three-dimensional landmarks, three-dimensional outlines and three-dimensional surfaces for representing morphological distinctions among three bivalve species belonging to the genus *Astarte* were undertaken to evaluate the potential of these data types to support automated species identification. While all three data types performed well (>70% identification accuracy) when compared to typical reproducibilities for human taxonomists using qualitative analyses, three-dimensional outlines and three-dimensional surfaces both achieved greater than 95 per cent identification accuracy. Between these two, three-dimensional surface analysis was clearly superior in terms of capturing the degree of interspecies-group distinctions evident to taxonomists.

Artificial neural net analyses of the set of raw *Astarte* images, along with high-resolution and lower resolution girded (= eigensurface) images of virtual specimens were also evaluated for interspecies-group distinctiveness. Results from these analyses indicated that, no matter how well processed, two-dimensional images cannot match the taxonomic information content of three-dimensional outlines and (especially) three-dimensional surfaces. However, previous investigations have suggested very high-quality neural net-based identifications can be made if training set sizes are sufficiently large (MacLeod et al., 2007b). Regardless, three-dimensional surface data have the greatest potential for supporting automated specimen recognition applications in systematics. The eigensurface approach represents a robust strategy for obtaining reliable, high-resolution, three-dimensional surface data that can be usefully exploited by either morphometric methods or ANNs. The potential of this approach spans the whole of morphology-based systematics and its development represents an important step towards completion of a comprehensive biological morphometrics toolkit.

Landmark and outline methods of shape characterization originally became popular not only because they were elegant and informative approaches to morphological analysis in their appropriate contexts, but also because they were the only approaches available to summarize morphological variation quantitatively. With the advent of three-dimensional digitizers and development of methods such as eigensurface analysis and ANNs that can extract both taxonomically representative and topologically comparable quantitative data from virtual specimens, the door has been opened for morphometricians and biologists to realize the benefits of a much more comprehensive approach to the study of the causes of morphological variation in organisms than was possible previously. But as before (see MacLeod, 2002a), the job of the morphometrician remains one of using these mathematical tools to help systematists–taxonomists do a better job. Similarly, the job of the taxonomist (in addition to being a taxonomist) should include assimilation of a general sense of what the morphometrician's current toolkit of methods can achieve, to use those tools and to work actively with morphometricians to improve them. Much error can be avoided if morphometricians and systematists–taxonomists both talk and listen to one another.

ACKNOWLEDGEMENTS

I would like to thank Quentin Wheeler for inviting me to contribute to his 'New Systematics' symposium at the 2005 Systematics Association Biennial Meeting in Cardiff and for being both patient and supportive during the manuscript's long gestation. Jonathan Krieger implemented my ideas for the three-dimensional gridding procedure developed as *Mathematica* code. Without Jon's *Mathematica* programming skills the manuscript would have had an even longer gestation. Mark O'Neill also stepped in to run the DAISY analyses. I thank both Jon and Mark for discussions of various aspects of the arguments presented herein. Natalie Dale-Skey provided invaluable assistance with collecting the three-dimensional data. Dave Polly is the co-developer of eigensurface analysis. Without his persistence, thoughtful collaboration and hard work it is doubtful the method would have been created. Finally, I thank Fred Bookstein and the organizers of the 2006 Vienna Morphofest for an invitation to attend that meeting. Aspects of the arguments presented herein–especially those concerning semilandmarks–were developed listening to several of the presentations at that meeting. Of course, none of the preceding are responsible for any errors of commission or omission in the text.

NOTES

1. Here, use of the term 'geometric' refers not to the sense that variables capturing aspects of shape are being analysed–as is the case with all forms of morphometrics–but rather (and obscurely) that these shape variables are being referenced to the geometry of the Kendall shape space (see Bookstein, 1991).
2. This is in addition to the facts that (1) quantitative methods of group recognition will also be needed to implement automated forms of species identification based on DNA barcodes and (2) the same generalized methods discussed in this chapter will very likely be used to accomplish any DNA-based, species-identification task.
3. In most cases the assessment of 'fatness' or 'thinness' will depend on object orientation, but this is a property of the object's geometry (e.g. unequal major and minor axes), not of its biology.
4. In principle, it is possible to test lower level, 'regional' homologies within a complex character state, but this is rarely done and would require access to complex evolutionary-developmental information, which would itself be subject to uncertainties for quantitative genetic traits at both higher and lower hierarchical levels.

5. These represent nothing more or less than objectively defined coordinate locations that bear some topological, anatomical, functional, etc. correspondence across all objects in the data-set. Such coordinates are often referred to as ‘homologous’ points, though they rarely correspond to biological homologies either in principle or in practice (see MacLeod, 1999, or Humphries, 2002, for further discussion).
6. The procedure described is referred to as generalized least squares (GLS) superposition. A number of variations on this theme exist (see Rohlf and Slice, 1990). These differ only in the details of the iteration strategy and/or the minimization metric (e.g. use of repeated median estimators for landmark positions).
7. In practice this sequence of steps (shape coordinates–relative warps–principal/partial warps) was often reordered to: shape coordinates–principal/partial warps–relative warps. However, since principal/partial warps simply represent a transformation of the original data and are decidedly unstable, most experienced practitioners recommend basing biological interpretations on the relative warps, which may be calculated directly from the shape coordinates.
8. If neural nets are trained using multimodal methods (e.g. simulated annealing or genetic algorithms), there is a far better chance of finding a global best answer.
9. This procedure is essentially equivalent to the adaptive outline-segment node spacing procedure used in extended eigenshape analysis; see MacLeod (1999).
10. Assuming that the training set samples are, on the whole, representative of the character and range of morphological variability in the species.

REFERENCES

Adams, D., Rohlf, F.J. and Slice, D.E. (2004) Geometric morphometrics: Ten years of progress following the ‘revolution’. *Italian Journal of Zoology,* 71: 5–16.

Barrow, E. and MacLeod, N. (in press) Shape variation in the mole dentary (Talpidae: Mammalia). *Biological Journal of the Linnean Society.*

Baylac, M. and Daufresne, T. (1996) Wing veination variability in *Monarthropalpus buxi* (Diptera, Cecidomyiidae) and the Quaternary coevolution of Box (*Buxus sempervirens* L.) and its midge: A geometrical morphometric analysis. In *Advances in Morphometrics* (eds L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor and D.E. Slice), Plenum Press, New York, pp. 285–301.

Becerra, J.M., Bello, E. and Garcia-Valdecasas, A. (1993) Building your own machine image system for morphometric analysis: A user point of view. In *Contributions to Morphometrics* (eds L.F. Marcus, E. Bello and A. García-Valdecasas), Monografias del Museo Nacional de Ciencias Naturales 8, Madrid, pp. 65–92.

Benson, R.H., Chapman, R.E. and Siegel, A.F. (1982) On the measurement of morphology and its change. *Paleobiology,* 8: 328–339.

Bishop, C.M. (1996) *Neural Networks for Pattern Recognition,* Oxford University Press, Oxford.

Blackith, R.E. and Reyment, R.A. (1971) *Multivariate morphometrics,* Academic Press, London.

Bogdanowicz, W. and Owen, R.D. (1996) Landmark-based size and shape analysis in systematics of the Plecotine bats. In *Advances in Morphometrics* (eds L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor and D.E. Slice), Plenum Press, New York, pp. 489–501.

Bookstein, F.L. (1978) *The Measurement of Biological Shape and Shape Change,* Springer, Berlin.

Bookstein, F.L. (1980) When one form is between two others: An application of biorthogonal analysis. *American Zoologist,* 20: 627–641.

Bookstein, F.L. (1986) Size and shape spaces for landmark data in two dimensions. *Statistical Science,* 1: 181–242.

Bookstein, F.L. (1989) Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11: 567–585.

Bookstein, F.L. (1990) Analytic methods: Introduction and overview. In *Proceedings of the Michigan Morphometrics Workshop* (eds F.J. Rohlf and F.L. Bookstein), The University of Michigan Museum of Zoology, Special Publication 2, Ann Arbor, MI, pp. 61–74.

Bookstein, F.L. (1991) *Morphometric Tools for Landmark Data: Geometry and Biology.* Cambridge University Press, Cambridge.

Bookstein, F.L. (1997) Landmark methods for forms without landmarks: Localizing group differences in outline shape. *Medical Image Analysis,* 1: 225–243.

Bookstein, F.L., Strauss, R.E., Humphries, J.M., Chernoff, B., Elder, R.L. and Smith, G.R. (1982) A comment on the uses of Fourier methods in systematics. *Systematic Zoology,* 31: 85–92.

Bookstein, F., Schäfer, K., Prossinger, H., Seidler, H., Fieder, M., Stringer, C., Weber, G.W., Arsuaga, J.-L., Slice, D., Rohlf, F.J., Recheis, W., Mariam, A.J. and Marcus, L.F. (1999) Comparing frontal cranial profiles in archaic and modern Homo by morphometric analysis. *The Anotomical Record,* 257: 217–224.

Burnaby, T.P. (1966) Growth invariant discriminant functions and generalized distances. *Biometrics,* 22: 96–110.

Cameron, S., Rubinoff, D. and Kipling, W. (2006) Who will actually use DNA barcoding and what will it cost? *Systematic Biology,* 55(5): 844–847.

Chu, K.H., Ho, H.Y., Li, C.P. and Chan, T.Y. (2003) Molecular phylogenetics of the mitten crab species in *Eriocheir, sensu lato* (Brachyura: Grapsidae). *Journal of Crustacean Zoology,* 23: 738–746.

Chu, K.H., Tong, J. and Chan, T.Y. (1999) Mitochondrial cytochrome oxidase I sequence divergence in some Chinese species of *Charybdis* (Crustacea: Decopoda: Portunidae). *Biochemical Systematics and Ecology,* 27: 461–468.

Colless, D.H. (1985) On ‘character’ and related terms. *Systematic Zoology,* 34: 229–233.

Culverhouse, P. (2007) Natural object categorization: Man vs. machine. In *Automated Taxon Recognition in Systematics: Theory, Approaches and Applications* (ed N. MacLeod), CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 25–45.

Culverhouse, P.F., Williams, R., Reguera, B., Herry, V. and González-Gils, S. (2003) Do experts make mistakes? *Marine Ecology Progress Series,* 247: 17–25.

Demeter, A., Rácz, G. and Csorba, G. (1996) Identification of house mice (*Mus musculus*) and mound-building mice (*M. spicilegus*), based on distance and landmark data. In *Advances in Morphometrics* (eds L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor and D.E. Slice), Plenum Press, New York, pp. 359–369.

Donoghue, M.J., Doyle, J.A., Gauthier, J., Kluge, A. and Rowe, T. (1989) The importance of fossils in phylogeny reconstruction. *Annual Review of Ecology and Systematics,* 20: 431–460.

Ebach, M.C. and Holdrege, C. (2005) DNA barcoding is no substitute for taxonomy. *Nature,* 434: 697.

Ehrlich, R., Pharr, R.B., Jr. and Healy-Williams, N. (1983) Comments on the validity of Fourier descriptors in systematics: A reply to Bookstein et al. *Systematic Zoology,* 31: 85–92.

Eldredge, N. and Cracraft, J. (1980) *Phylogenetic Patterns and the Evolutionary Process,* Columbia University Press, New York.

Firstrup, K. (1992) Character: Current usages. In *Keywords in Evolutionary Biology* (eds E.F. Kellerand and E.A. Lloyd), Harvard University Press, Cambridge, MA, pp. 45–51.

Fisher, R.A. (1936) The utilization of multiple measurements in taxonomic problems. *Annals of Eugenics,* 7: 179–188.

Foote, M. (1989) Perimeter-based Fourier analysis: A new morphometric method applied to the trilobite cranidium. *Journal of Paleontology,* 63: 880–885.

Foote, M. (1991) Morphologic patterns of diversification: Examples from trilobites. *Palaeontology,* 34: 461–485.

Foote, M. (1993) Discordance and concordance between morphological and taxonomic diversity. *Paleobiology,* 19: 185–204.

Foote, M. (1996) Models of morphological diversification. In *Evolutionary Paleobiology* (eds D. Jablonski, D.H. Erwin and J. Lipps), University of Chicago Press, Chicago, pp. 63–86.

Gayon, J. (2000) History of the concept of allometry. *American Zoologist,* 40(5): 748–758.

Godfray, H.C.J. (2002) Challenges for taxonomy. *Nature,* 417: 17–19.

Goodall, C.R. (1991) Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society, Series B,* 53: 285–339.

Gower, J.C. (1976) Growth-free canonical variates and generalized inverses. *Bulletin of the Geological Institute, Uppsala University,* 7: 1–10.

Gubányi, A. (1996) Morphometric analysis of microscopic hooks of taeniid tapeworms (Cestoda, Taeniidae). In *Advances in Morphometrics* (eds L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor and D.E. Slice), Plenum Press, New York, pp. 503–510.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B,* 270: 313–322.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. and Hallwacks, W. (2005) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astaptes fulgerator. Proceedings of the National Academy of Sciences,* 101: 14812–14817.

Hickerson, M.J., Meyer, C.P. and Moritz, C. (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology,* 55(5): 729–739.

Hugot, J.P. and Baylac, M. (1996) Comparative landmark analysis of various Oxyurudae parasites of primates and rodents. In *Advances in Morphometrics* (eds L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor and D.E. Slice), Plenum Press, New York, pp. 463–478.

Hull, D.L. (1988) *Science as a Process: An Evolutionary Account of the Social and Conceptual Developments in Science,* University of Chicago Press, Chicago.

Humphries, C.J. (2002) Homology, characters and continuous variables. In *Morphology, Shape and Phylogeny* (eds N. MacLeod and P.L. Forey), Taylor & Francis, London, pp. 8–26.

Huxley, J.S. (1932) *Problems of relative growth,* Methuan & Co., London.

Huxley, J.S. and Tessier, G. (1936) Terminology of relative growth. *Nature,* 137: 780–781.

Janzen, D.H. (2004) Now is the time. *Philosophical Transactions of the Royal Society of London, Series B,* 359: 731–732.

Jolicouer, P. (1963) The multivariate generalization of the allometry equation. *Biometrics,* 19: 497–499.

Jolicouer, P. and Mosimann, J.E. (1960) Size and shape variation in the painted turtle, a principal component analysis. *Growth,* 24: 339–354.

Kaesler, R.L. (1993) A window of opportunity: Peering into a new century of paleontology. *Journal of Paleontology,* 67: 329–333.

Kendall, D. (1977) The diffusion of shape. *Advances in Applied Probability,* 9: 428–430.

Kendall, D.G. (1984) Shape manifolds, procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society,* 16: 81–121.

Kitching, I.J., Forey, P.L., Humphries, C.J. and Williams, D.M. (1998) *Cladistics: The Theory and Practice of Parsimony Analysis,* 2nd ed., Oxford University Press, Oxford.

Klingenberg, C.P. and Monteiro, A.R. (2005) Distances and directions in multidimensional shape spaces: Implications for morphometric applications. *Systematic Biology,* 54: 678–688.

Lang, R. (2007) Neural networks in brief. In *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications* (ed N. MacLeod), CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 47–68.

Lohmann, G.P. (1983) Eigenshape analysis of microfossils: A general morphometric method for describing changes in shape. *Mathematical Geology,* 15: 659–672.

MacLeod, N. (1990) Digital images and automated image analysis systems. In *Proceedings of the Michigan Morphometrics Workshop* (eds F.J. Rohlf and F.L. Bookstein), The University of Michigan Museum of Zoology, Special Publication 2, Ann Arbor, MI, pp. 21–35.

MacLeod, N. (1998) Impacts and marine invertebrate extinctions. In *Meteorites: Flux with Time and Impact Effects* (eds M.M. Grady, R. Hutchinson, G.J.H. McCall and D.A. Rotherby), Geological Society of London, London, pp. 217–246.

MacLeod, N. (1999) Generalizing and extending the eigenshape method of shape visualization and analysis. *Paleobiology,* 25(1): 107–138.

MacLeod, N. (2001) Landmarks, localization, and the use of morphometrics in phylogenetic analysis. In *Fossils, Phylogeny, and Form: An Analytical Approach* (eds G. Edgecombe, J. Adrain and B. Lieberman), Kluwer Academic/Plenum, New York, pp. 197–233.

MacLeod, N. (2002a) Phylogenetic signals in morphometric data. In *Morphology, Shape and Phylogeny* (eds N. MacLeod and P.L. Forey), Taylor & Francis, London, pp. 100–138.

MacLeod, N. (2002b) Geometric morphometrics and geological form-classification systems. *Earth-Science Reviews,* 59(2002): 27–47.

MacLeod, N. (2005) Shape models as a basis for morphological analysis in palaeobiological systematics: Dicotyledenous leaf physiography. *Bulletins of American Paleontology,* 369: 219–238.

MacLeod, N. (2007a) Groups I. *Palaeontological Association Newsletter,* 64: 35–45.

MacLeod, N. (2007b) Groups II. *Palaeontological Association Newsletter,* 65: 36–49.

MacLeod, N., Ed (2007c) *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications,* Taylor & Francis, London.

MacLeod, N. and Polly, P.D. (2005) A new eigenshape-based morphometric method for analyzing three-dimensional patterns of shape variation for surfaces and objects. *Palaeontological Association Newsletter,* 60: 23.

MacLeod, N. and Rose, K.D. (1993) Inferring locomotor behavior in Paleogene mammals via eigenshape analysis. *American Journal of Science,* 293-A: 300–355.

MacLeod, N., O'Neill, M.A. and Walsh, S.A. (2007a) A comparison between morphometric and artificial neural net approaches to the automated species-recognition problem in systematics. In *Biodiversity Databases: From Cottage Industry to Industrial Network* (eds G. Curry and C. Humphries), Taylor & Francis, London, pp. 37–62.

MacLeod, N., O'Neill, M. and Walsh, A.S. (2007b) Automated tools for the identification of taxa from morphological data: Face recognition in wasps. In *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications* (ed N. MacLeod), Taylor and Francis, London, pp. 153–188.

Marcus, L.F., Bello, E. and García-Valdecasas, A. (1993) *Contributions to Morphometrics,* Museo Nacional de Ciencias Naturales 8, Madrid.

Marcus, L.F., Corti, M., Loy, A., Naylor, G.J.P. and Slice, D.E., eds (1996) *Advances in Morphometrics,* Plenum Press, New York.

Mardia, K.V. and Dryden, I. (1989a) The statistical analysis of shape data. *Biometrika,* 76: 271–282.

Mardia, K.V. and Dryden, I.L. (1989b) Shape distributions for landmark data. *Advances in Applied Probability,* 21: 742–755.

Meier, R., Shiyang, K., Vaidya, G. and Ng, P.K.L. (2006) DNA barcoding an taxonomy in Diptera: A tale of high interspecific variability and low identification success. *Systematic Biology,* 55(5): 715–728.

O'Neill, M. (2007) *DAISY*: A practical tool for semi-automated species identification. In *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications* (ed N. MacLeod), CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 101–114.

Pimentel, R.A. (1979) *Morphometrics: The Multivariate Analysis of Biological Data,* Kendall/Hunt, Dubuque, IA.

Polly, P.D. and MacLeod, N. (2005) Characterization and comparison of three-dimensional shapes using eigensurface analysis: Locomotion in Tertiary Carnivora. *Journal of Vertebrate Paleontology,* 26(Supplement to No. 3): 111A.

Polly, P.D. and MacLeod, M. (in press) Locomotor in fossil Carnivora: An application of the eigensurface method for morphometric analysis of three-dimensional surfaces. *Palaeontologia Electronica,* 10(1).

de Queiroz, K. and Donoghue, M.J. (1988) Phylogenetic systematics and the species problem. *Cladistics,* 4: 317–338.

de Queiroz, K. and Donoghue, M.J. (1990) Phylogenetic systematics and species revisited. *Cladistics,* 6: 83–90.

Reyment, R.A. (1971) *Quantitative palaeoecology,* Elsevier, Amsterdam.

Reyment, R.A. (1980) *Morphometric methods in biostratigraphy,* Academic Press, London.

Reyment, R.A. (1991) *Multidimensional paleobiology,* Pergamon Press, Oxford.

Reyment, R.A., Blackith, R.E. and Campbell, N.A. (1984) *Multivariate morphometrics,* 2nd ed., Academic Press, London.

Rohlf, F.J. (1993) Relative warp analysis and an example of its application to mosquito wings. In *Contributions to Morphometrics* (eds L.F. Marcus, E. Bello and A. García-Valdecasas), Museo Nacional de Ciencias Naturales 8, Madrid, pp. 131–160.

Rohlf, F.J. and Bookstein, F.L., eds (1990) *Proceedings of the Michigan Morphometrics Workshop.* The University of Michigan Museum of Zoology Special Publication 2, Ann Arbor.

Rohlf, F.J. and Slice, D. (1990) Extensions of the Procrustes method for optimal superposition of landmarks. *Systematic Zoology,* 39: 40–59.

Rudwick, M.J.S. (1972) *The Meaning of Fossils: Episodes in the History of Palaeontology,* MacDonald, London.

Sampson, P.D., Bookstein, F.L., Sheehan, F.H. and Bolson, E.L. (1996) Eigenshape analysis of left ventricular outlines from contrast ventriculograms. In *Advances in Morphometrics* (eds L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor and D.E. Slice), Plenum Press, New York, pp. 211–234.

Scotland, R. and Pennington, R.T. (2000) *Homology and Systematics: Coding Characters for Phylogenetic Analysis,* Taylor & Francis, London.

Shih, H.-T., Ng, P.K.L. and Chang, H.-W. (2004) The systematics of the genus *Geothelphusa* (Crustacea: Decapoda: Brachyuridae: Potamidae) from southern Taiwan: A molecular appraisal. *Zoological Studies,* 43: 561–570.

Siegel, A.F. and Benson, R.H. (1982) A robust comparison of biological shapes. *Biometrics,* 38: 341–350.

Smith, A.B. (1994) *Systematics and the Fossil Record: Documenting Evolutionary Patterns,* Blackwell, London.

Sneath, P.H.A. (1967) Trend surface analysis of transformation grids. *Journal of Zoology,* 151: 65–122.

Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy: The Principles and Practice of Numerical Classification,* W.H. Freeman, San Francisco.

Sokal, R.R. (1966) Numerical taxonomy. *Scientific American,* 215(6): 106–116.

Sokal, R.R. and Rohlf, F.J. (1966) Random scanning of taxonomic characters. *Nature,* 210: 461–462.

Sokal, R.R. and Sneath, P.A. (1963) *Principles of numerical taxonomy,* W.H. Freeman, San Francisco.

Strauss, R.E. and Bookstein, F.L. (1982) The truss: Body form reconstruction in morphometrics. *Systematic Zoology,* 31: 113–135.

Sun, H.Y., Zhou, K. and Yang, X.J. (2003) Phylogenetic relationships of the mitten crabs inferred from mitochondrial 16S rDNA partial sequences (Crustacean, Decapoda). *Acta Zoologica Sinica,* 49: 592–599.

Tang, B., Zhou, K., Song, D., Yang, G. and Dai, A. (2003) Molecular systematics of the Asian mitten crabs, genus *Eriocheir* (Crustacea: Brachyura). *Molecular Phylogenetics and Evolution,* 29: 309–316.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2003) A plea for DNA taxonomy. *Trends in Ecology and Evolution,* 18: 70–74.

Thompson, D.W. (1917) *On Growth and Form,* Cambridge University Press, Cambridge, England.

Todd, J. (2002) Bivalves. In *PaleoBase: Macrofossils 2* (ed N. MacLeod), Blackwell Science and The Natural History Museum, London.

Weeks, P.J.D., Gauld, I.D., Gaston, K.J. and O'Neill, M.A. (1997) Automating the identification of insects: A new solution to an old problem. *Bulletin of Entomological Research,* 87: 203–211.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D. (1999a) Automating insect identification: exploring limitations of a prototype system. *Journal of Applied Entomology,* 123: 1–8.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D. (1999b) Species identification of wasps using principal component associative memories. *Image and Vision Computing,* 17: 861–866.

Wheeler, Q.D. (2003) Transforming taxonomy. *The Systematist,* No. 22: 3–5.

Wheeler, Q.D. (2005) Losing the plot: DNA 'barcodes' and taxonomy. *Cladistics,* 21(2005): 405–407.

Wheeler, Q.D. (2007) Digital innovation and taxonomy's finest hour. In *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications* (ed N. MacLeod), CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 9–23.

Wheeler, Q. and Meier, R. (2000) *Species Concepts and Phylogenetic Theory,* Columbia University Press, New York.

Wheeler, Q.D. and Nixon, K.C. (1990) Another way of looking at the species problem: A reply to de Queiroz and Donoghue. *Cladistics,* 6: 77–81.

Wheeler, Q.D., Raven, P.H. and Wilson, E.O. (2004) Taxonomy, impediment or expedient? *Science,* 303: 285.

Wood, S.V. (1850–1856) A monograph of the Crag mollusca with descriptions of shells from the Upper Tertiaries of the British Isles. *Palaeontographical Society Monographs,* Vol. 2, 1863 pp.

Zahn, C.T. and Roskies, R.Z. (1972) Fourier descriptors for plane closed curves. *IEEE Transactions, Computers,* C-21: 269–281.

Zelditch, M.L., Fink, W.L. and Swiderski, D.L. (1995) Morphometrics, homology, and phylogenetics: Quantified characters as synapomorphies. *Systematic Biology,* 44: 179–189.

Zelditch, M.L., Swiderski, D.L., Sheets, H.D. and Fink, W.L. (2004) *Geometric Morphometrics for Biologists: A Primer,* Elsevier/Academic Press, Amsterdam.

Zhi, L., Karesh, W.B., Janczewski, D.N., Frazier-Taylor, H., Sajuthi, D., Gombek, F., Andau, M., Martenson, J.S. and O'Brien, S.J. (1996) Genomic differentiation among natural populations of orangutan (*Pongo pygmaeus*). *Current Biology,* 6: 1326–1336.

10 Taxonomic Shock and Awe

Quentin D. Wheeler

CONTENTS

> Since before Sun Tzu and the earliest chroniclers of war recorded their observations, strategists and generals have been tantalized and confounded by the elusive goal of destroying the adversary's will to resist before, during, and after battle.
>
> **(Harlan Ullman and James Wade)**

> An army of principles can penetrate where an army of soldiers cannot.
>
> **(Thomas Paine)**

INTRODUCTION

The greatest contributions of taxonomy to science and humanity are yet to come. Against formidable odds and with minimal funding, equipment, infrastructure, organization and encouragement, taxonomists have discovered, described and classified nearly 1.8 million species. While increasing attention is being paid to making this substantial amount of accumulated taxonomic information more easily accessible, comparatively little attention has been paid to opening access to the research resources required by taxonomists themselves. Benefits associated with ease of access to museum records (e.g. Global Biodiversity Information Facility) or 'known' species (e.g. Encyclopedia of Life) are seriously restricted when such information is untested for validity or is simply unavailable, as is the case for three-quarters or more of the species on Earth. We act as if taxonomy is done but nothing could be farther from the truth. Even in comparatively well-known areas of the world such as Europe or North America, much remains to be done in both describing new species and testing putatively 'known' ones. For most areas of the world, including those with the largest numbers of species, the situation is far worse and the process of species exploration has scarcely begun.

Taxonomists could work at rates orders of magnitude faster were they to adopt more efficient and collaborative modes of work, engineer more appropriate and efficient instruments and tools and construct a domain-specific cyberinfrastructure specially designed to meet the unique needs of taxonomy. Such an infrastructure would link all the resources needed by a taxonomist in his or her routine work, including museum specimens, archival literature, specimen databases, high-end instrumentation and taxon experts. Many components of this research platform exist or are under active construction, including new modes of work. For example, an international 'team' approach to taxonomy is emerging (see Knapp, this volume). Thus, calls to replace the information-rich sciences of comparative morphology, paleontology and ontogeny with simpler, less informative data sources are premature and unnecessary. The traditional goal of an integrated taxonomy informed by all relevant sources of evidence–what Hennig termed 'holomorphology'–is now more easily and fully achievable than ever. Virtually every constraint of the past is being overcome or can easily be overcome (Wheeler, 2004; Wheeler et al., 2004; Page et al., 2005) and a glimpse of a cyber-enabled massive-scale taxonomic enterprise has been seen in the proposed Legacy Infrastructure Network for Natural Environments (LINNE) project (Page et al., 2005).

The US National Science Foundation has funded a series of Planetary Biodiversity Inventory (PBI) projects, each characterized by tackling a large monophyletic group through focused and coordinated international teams of experts, students and museums. Thousands of species are discovered and described or corroborated and redescribed in a single 5-year funded period. The success of the PBI model takes advantage of the comparative method of taxonomy and of efficiencies of teamwork and dispels the myth that species description and testing is necessarily slow. These projects have labored largely with existing instruments and technology. Imagine how much more efficient such projects could be were the vision of a taxonomy cyberinfrastructure realized, such as the proposed LINNE project (see Page et al., 2005).

This chapter is titled after a military report by Ullman and Wade (1996) that discussed the 'shock and awe' strategy of warfare. The overarching concept is an attack that is so massive and overwhelming that the enemy is instantly demoralized and simply unable to respond. This chapter is a call to arms for taxonomists and for museums and herbaria that are the most vital infrastructure for taxonomy.

By working together and speaking with a single voice, the taxonomic community can build a LINNE-like taxonomy-specific cyberinfrastructure that gives to taxonomists the capacity to launch a species exploration 'shock and awe' campaign capable of leaving the opposition who deny to taxonomy its proper status and funding speechless.

THE 2020 CHALLENGE

This taxonomic shock and awe campaign will be the most ambitious exploration of species and reconstruction of evolutionary history ever conceived. Properly equipped and funded, the taxonomic and museum communities can advance our knowledge of species more rapidly over decades than has been possible previously over centuries.

I challenge the community to focus first on building a virtual research 'instrument' … a kind of distributed 'species observatory' by the year 2020. This species observatory will seamlessly link the estimated three billion specimens in the world's natural history museums and herbaria with researchers and students of taxonomy; all published literature since 1758 in digital form; digitized records associated with specimens in museums; digital and remotely accessible instruments to observe, measure, photograph, manipulate and dissect specimens; an advanced cladistic analysis and descriptive taxonomic software toolkit; and a range of other 'tools' required by various taxon specialists and also of use in the education of taxonomists and to users of taxonomic information.

Where we see today hundreds of museums with local, provincial or national foci, tomorrow we will see a single world 'museum' linked in cyberspace. This museum represents a treasure of humanity and houses our sum total documentation of life on our planet. Because taxonomists deal with planetary-scale hypotheses spanning all geographical and temporal boundaries, taxonomists require such a world collection and worldwide research platform. Each museum gives up a little autonomy but in exchange avoids redundancy, enjoys specialization and gains open access to every other collection on Earth.

This is a proposition of enormous audacity. But anything short of completeness or comprehensiveness is simply unacceptable to the screaming need of the world for complete and reliable information about what species exist, what properties they have, how they are related and where they are distributed. This is the only basis on which taxonomy will realize its enormous and frustrated potential. With such an infrastructure, taxonomy will open to humanity the full grandeur and possibilities that are life on our planet.

Beginning now and building upon the rapid advances being made around the world in digitization of museum data, archival literature and species information, this dream is very much in reach. The foundations of cyberinfrastructure are being conceived and designed by leading computer scientists (Atkins et al., 2003); thus, we can focus on a taxonomy and museum infrastructure that simply plugs in and takes advantage of it. While fully supporting the ambitious goal of putting every known species on the web (e.g. the Encyclopedia of Life project), taxonomists know that the usefulness and reliability of those 1.8 million pages will be severely limited unless taxonomy is able to far more rapidly and efficiently test and corroborate those species and discover, describe and make available to such projects the other eight or more millions of species alive today.

Everyone assumes that taxonomic information will be available and that taxonomists will simply create it while being unfunded and ignored. The incredibly impressive excellence of taxonomic theory is being undermined by efforts to supplant it with rapid but vastly inferior alternatives such as DNA taxonomy and DNA barcoding. Let me be clear. I am not indicting DNA or other molecular evidence. It has earned a prominent position among the sources of evidence of taxonomy. It is incredibly valuable in both phylogenetic studies and in species identifications and in associating semaphoronts within species (Miller et al., 2006). However, a DNA taxonomy is a sad alternative to integrative taxonomy and DNA barcoding is, at best, a good identification tool but a very limited species discovery tool (Meyer and Paulay, 2005; Prendini, 2005; Wheeler, 2005; Will et al., 2005; Meier et al., 2006; Little and Stevenson, 2007).

At the same time that the taxonomy cyberinfrastructure is being built over the next dozen years–let us call it the 2020 project both because it can be completed by the year 2020 and because this observatory will give us 20/20 vision of the world's species–two kinds of communities should be self-organizing and undertaking massively scaled species exploration projects. Species exploration here may be understood to include both the discovery and description of new species and the continual critical testing of existing species hypotheses.

One is a community of museums who, due to limited resources, need to work together to coordinate worldwide species exploration and worldwide collection growth and development to make the best use of funds and resources and expertise. Exploring the biosphere of Earth is not unlike sharing seismic data with colleagues in order to understand the lithosphere: It demands international cooperation and sharing of data. No museum can be complete. No taxonomist can work in complete isolation. No taxon can be known without global sharing of data and specimens.

The other is a community of taxon experts who pull together to assemble, test, maintain and expand our sum total knowledge of a taxon. The concepts of a taxon knowledge bank and of taxon knowledge communities are discussed later. In this context, let us say that such communities should be organized and fully functional by 2020. That means every essential collection, expert, published paper, digitized image, or recorded data should be compiled and accessible to that research community. There should be hundreds of such communities in place and active by 2020. In fact, just as the Encyclopedia of Life, if successful, will have nearly two million species pages on the web delivering what is known by then, taxonomists should have all the background information associated with that knowledge similarly compiled into the kind of taxon knowledge banks described in this chapter.

THE CALL TO ARMS

I am as guilty as any taxonomist and perhaps guiltier than most in whining about the state of taxonomy. It is difficult for anyone who appreciates the rigorous theoretical basis for phylogenetic species (Nixon and Wheeler, 1992; Wheeler and Platnick, 2000) and phylogenetic systematics (Hennig, 1966; Nelson and Platnick, 1981;

Schuh, 2000) to not feel despair and anger at how taxonomy has been marginalized and underfunded for decades (Wheeler, 1995a, 2004). It is a simple reality that whiners seldom gain what they need. Need alone is not much of a motivator for donors or supporters. Everyone would prefer to back a winner. Thus, taxonomy needs a dramatic change in tactics and to take advantage of existing and reachable technologies to mount an aggressive, focused and awesome-scaled campaign of species discovery and testing. The ultimate goal should be to overwhelm the skeptics, detractors and enemies of taxonomy with so much evidence of such high quality and obvious relevance to science and society that its importance can no longer be denied.

Having had the privilege of serving as a division director at the US National Science Foundation, I saw on several occasions the awesome power of a community of scientists with the political savvy to speak with a unified voice. No public funding source can afford to ignore the demands of any sizable community of scientists who represent numerous congressional districts. It is also true that theoretical and practical advances in science must be won by the traditional method: by using publications and public lectures and direct interactions with colleagues to change minds. It is the scientific community that sets its standards, both good and bad. Taxonomists are so busy fighting one another over the fine-grained arguments necessary for good science that they fail to see the possibilities for working together to secure additional support for taxonomy as a whole. This is nothing more than the ancient divide-and-conquer strategy, in this case self imposed. All taxonomists need is to have enough faith in science to realize that even if silly or wrongheaded proposals enjoy temporary support (the Phylocode and DNA barcoding leap to mind in this context) that they will ultimately be discredited and abandoned (for responses to the Phylocode, see Dominguez and Wheeler, 1992; Carpenter, 2003; Nixon et al., 2003; Keller et al., 2003; and to DNA barcoding, see Meyer and Paulay, 2005; Prendini, 2005; Wheeler, 2005; Will et al., 2005; Ebach and Holdrege, 2005; Meier et al., 2006; Little and Stevenson, 2007). In fact, it is sometimes necessary that some funding be directed to such intellectual cul-de-sacs so that enough empirical data accumulate to demonstrate that they are bad ideas.

Taking such a broad view of taxonomy and believing in the essentiality of taxonomy as a counterbalance to general (or experimental or functional) biology (Nelson and Platnick, 1981), the taxonomic and museum communities need to come together and demand investments in infrastructure from growing and developing collections to delivering the kind of cyberinfrastructure envisaged by LINNE to building a new generation of instruments and educating a new generation of taxon specialists. If we make such a univocal demand and define an ambitious but achievable agenda for the next 2 years, I believe that we can and will succeed.

What does a New Taxonomy capable of executing a shock and awe campaign look like? What will be required to make both a New Taxonomy and a successful campaign successful?

ATTRIBUTES OF THE NEW TAXONOMY

The New Taxonomy must first reassert its status as an independent science. Taxonomic knowledge and information are necessary for credible biology, but this does not mean that taxonomy exists or that the best taxonomy is done to meet the needs of biology. To the contrary, an active, independent taxonomic research enterprise is the best source for the highest quality results. I do not mean to take away from the importance of applied taxonomy but rather to emphasize that taxonomy has its own questions, epistemology and needs and that these are never met fully through approaches that do not respect taxonomy as a science.

Such applied outputs are more than sufficient reason to fund taxonomy fully: corroborated species, character and clade hypotheses; ability to accurately identify species; a precise and informative language of biological diversity in the form of Linnaean names associated with phylogenetic classifications; predictions possible with such a classification; and so forth. The best hypotheses, however, emerge from taxonomy done for its own sake. Rather than focusing on a few economically important species, a taxonomic study focuses on an entire clade and rigorously reviews all its constituent species. Through such comprehensively comparative taxonomy come the best and most fully tested hypotheses about species, characters and monophyletic groups (higher taxa).

Assuming that the New Taxonomy will be designed to meet the needs of taxonomists themselves and thereby result in the most (and most reliable) information for other branches of biology, let us examine some of the attributes of this New Taxonomy.

Planetary-Scale Projects

Taxonomy done well involves the comprehensive review of each and every species in a monophyletic group without regard to their geographical or ecological distributions or restricted by any arbitrary scale of time. Answers to taxonomy's 'big questions' demand such unrestricted research projects (Cracraft, 2002). In order to practice taxonomy in such an unrestricted way, it is imperative that access be opened to the full range of research resources required: specimens, specimen data, all relevant literature, instruments and colleagues.

Accelerated Species Testing

Species are described, but those descriptions are hypotheses that make explicit predictions about the distribution of characters (Wheeler and Platnick, 2000). Unless newly collected specimens and newly discovered characters are used to repeatedly challenge such hypotheses, they do not reflect our best estimate of nature. Tests of single species hypotheses are costly and inefficient because they involve examination of numerous related species with which the target species might be confused. Similarly, collecting specimens is most efficient when all related species at field sites are collected simultaneously. A species by species approach is inherently inefficient and it belies the strength of the comparative method of taxonomy.

The traditional high-throughput species testing tool was the taxonomic revision or monograph. In such studies all related species, all accumulated and available specimens in museums and all characters discovered since the preceding revision are reviewed and compared comprehensively and in detail. This adds both efficiency and unique insights by virtue of assessing all characters and states at once. Such studies invariably include the addition of new species or characters in addition to critically testing previously documented species and characters. This is a fundamentally sound model for species testing. The only limitation is that the comprehensive nature of such studies combined with a paucity of taxonomists has meant that such corroboration and refutation of species has taken place only infrequently. In entomology, for example, it is common for the gap between such revisions to be 50–100 years. During that interval it becomes increasingly difficult to identify species, determine which are still believable or have confidence in the distribution of attributes among them.

The New Taxonomy will be built upon a taxon knowledge community model (see next section). This means that species are continuously tested as soon as new specimens or characters become available. In the old revision model the same tests took place, but often over generations. In the New Taxonomy a character or species hypothesis that is incorrect may well be falsified and rejected within days or weeks instead. This is because as soon as a new species is hypothesized (described) the specimens in the type series are accessible to colleagues who may compare them with new evidence or specimens at their disposal.

Taxon Knowledge Communities

Information scientists talk about 'knowledge communities' as people with similar interests and expertise connected through cyberinfrastructure such that they may interact, collaborate and even compete with one another on a much accelerated time scale. All the experts in the world will in the New Taxonomy be linked into such knowledge communities (Figure 10.1). Those experts will collectively manage all the hypotheses, data and information associated with the taxon for which they are the experts. They will work in a shared cyber space. Within that space will exist the sum total of human knowledge of the group: every character hypothesis, every species description, every specimen in every museum, every database of observations and measurements, every published account of any of its species. This collective knowledge will be maintained in a virtual, dynamic, constantly advancing electronic 'monograph' or "knowledge bank" (Figure 10.2).

Some researchers will collaborate in PBI style teams to advance our knowledge of some subtaxa. Others will choose to work alone, but will draw from this shared knowledge base as they need and 'publish' their conclusions as a part of that same knowledge base. Sophisticated software will make it possible for the user to follow any of several competing hypotheses presented by any number of experts, to reanalyse phylogenetic relationships using their own parameter specifications and to generate designer 'publications' (delivered in printed or electronic form) to meet their particular needs. Examples will include revisions, diagnostic or illustrated keys, distribution maps, host lists, check lists for geographical places and so forth.

Taxonomy-Specific Cyberinfrastructure

The LINNE project has begun to characterize what a taxonomy-specific cyberinfrastructure must look like (Page et al., 2005). Some have misconstrued the LINNE model as little more than a data access scheme, like the Global Biodiversity Information Facility. GBIF is critically important but is ultimately only a reliable source

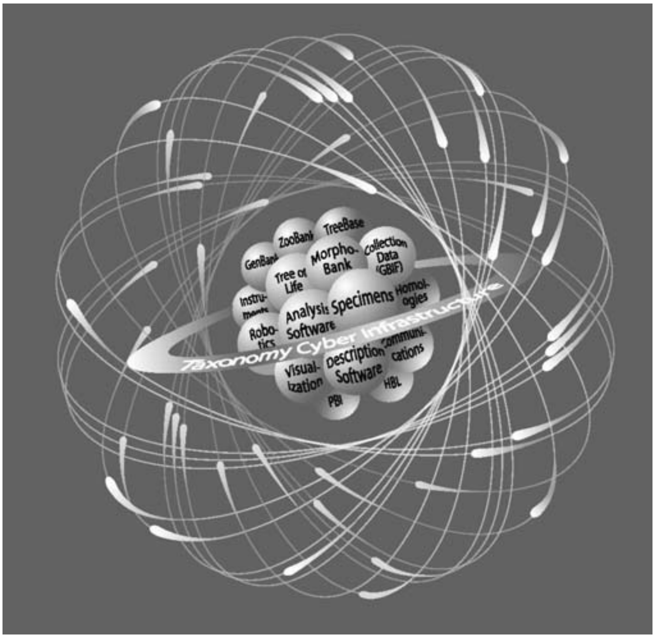


FIGURE 10.1 A graphic conception of a taxon knowledge community. Taxonomists, represented by the orbiting 'electrons', work individually or converge temporarily to work as members of teams to make rapid progress on one subtaxon. The world's experts on a taxon all work in this virtual taxon space, drawing from a 'nucleus' of resources through a foundation of cyberinfrastructure. That nucleus includes museum specimens, analytical software, remotely accessible digital instruments, databases, literature, ZooBank, MorphoBank, robotics associated with specimen collection, preparation and study, visualization tools for morphological characters and phylogenies, etc. Graphic courtesy of Frances Fawcett.

of information when fundamental taxonomy is done. Museum collections rely upon taxonomic revisions to constantly improve species concepts and assure that species identifications are correct. In the absence of such revisionary studies, museum data erodes in quality and ultimately becomes unreliable (see, for example, Meier and Dikow, 2004).

Aspects of the taxonomy cyberinfrastructure have been described elsewhere (Wheeler, 2004, 2007; Page et al., 2005) but the full breadth of possibilities is yet to be imagined. Among the minimal capabilities are the following. Remote access to specimens can allow them to be manipulated and photographed. This could easily be expanded to include remote dissections through devices similar to the Da Vinci surgical instrument. As experts create digital images of specimens, particularly types, these images will populate an ever expanding digital library. That library includes or is a part of public repositories of visual documentation of morphology such as the MorphBank and MorphoBank projects. Efforts are well under way to digitize archival literature; this is important to all fields of science, but vital to taxonomy where any species named after January 1, 1758, must be taken into account for complete revisionary work. Video communication tools are essential. Experts around

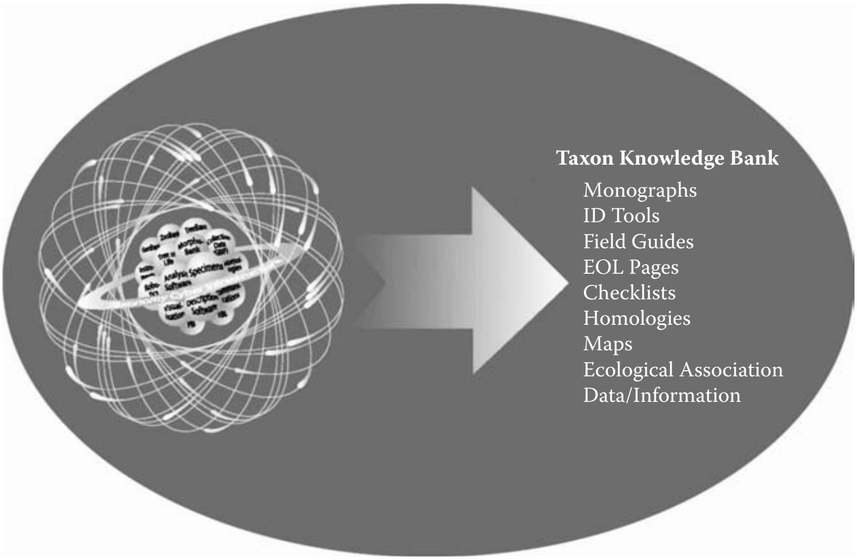


FIGURE 10.2 From a taxon knowledge community's collective research activities comes a virtual 'taxon knowledge bank'. This knowledge base is the sum total of what we know of a taxon. Taxonomists withdraw the resources needed to do taxonomy and deposit their results and hypotheses. From this taxon knowledge bank users of taxonomic information may withdraw what they need in any form in which they may need it. Common examples might include monographs, identification tools, field guides, Encyclopedia of Life web pages, checklists, maps and so forth. Graphic courtesy of Frances Fawcett.

the world need to collaborate in cyberspace. These collaborative communication software tools will ultimately include seeing a specimen together in real time along with relevant literature, images, etc.; seeing one another either in real time or in delayed communications to account for time zones, rather like an elaborate chess game played across continents; and communicating through software that translates to eliminate language barriers between collaborating taxon experts.

Ultimately, as the taxonomy cyberinfrastructure grows, it will become the world's first *species observatory.* Because revisionary taxonomy requires comparing thousands of specimens from scores of museums in as many countries as well as hundreds of characters recorded from specimens representing geographically and genetically diverse populations, it has been nearly impossible in the past to make all the necessary direct comparisons. In cybertaxonomy, it will be possible to compare any two specimens from any two museums anywhere in the world side by side in high-resolution digital splendor. The growth of cybertaxonomy may proceed rapidly if a major infrastructure project like LINNE is funded. Even if so, it will continue to grow through successive waves of innovation, each overcoming something previously accepted as a limitation. There is no longer justification to live with limited access to anything that taxonomists require to do their work, full stop. These advances are easily envisioned at every step of what we know is good taxonomy–from collecting and preparing specimens to studying specimens, recording and analysing characters, accessing research resources and disseminating knowledge and discoveries.

Electronic Apprenticeships

Teaching taxonomy will no longer require that the master, the student or the specimens be in the same institution or city for a class. A world expert in London may be manipulating a rare specimen in Paris while teaching a class on a taxon to students in Brazil, Kenya and New Zealand.

One-Stop Shopping for Names

A recent proposal to require registration of names in zoology (Polaszek et al., 2004) should be extended to botany as well. In the midst of a biodiversity crisis, it is absurd to continue to allow new species to be described in obscure journals that it may take years to discover. This registration scheme should threaten no one as it imposes no restrictions on where such names are published or any form of censorship, merely the notification of a single database of the existence of new species descriptions and names. We continue to play catch up with 250 years of names; there is no justification for adding even a single additional name to this chaos.

A Species Identification Arsenal

There is a rapidly expanding arsenal of tools that make species identification possible. For tissue samples or fragments of specimens or diverse life stages DNA signatures can be used to make identifications by comparing appropriately variable and reliable genes with a library of known and corroborated species (Miller et al., 2005; Little and Stevenson, 2007). For specimens in the field, handheld devices can download picture-rich identification software that involves answering a series of questions or matching the image to the specimen in order to complete an identification. Computer automated image comparisons also make it possible to mechanize morphometric-based identifications (see MacLeod, 2007). Current efforts to deny funding to comparative morphology, behavior and other studies in favor of molecular data alone are offered by non-taxonomists unaware of the theories or history of taxonomy or, it seems, of the obvious benefits of multiple data sources (Will et al., 2005; Desalle, 2006).

Unit Species Concept

The New Taxonomy should include an immediate and intensive period of debate to arrive at a unit species concept. Unless new theoretical arguments emerge, there is no reason to believe that multiple concepts are necessary. To the contrary, the phylogenetic species concept (PSC) appears to fulfill this role (Nixon and Wheeler, 1992; Wheeler, 1997; Wheeler and Platnick, 2000). The PSC is only a reflection of patterns of character distribution. Much like a cladogram, this is a necessary first step that may be followed by questions about evolutionary processes. The advantage is that as a pattern the PSC is consistent with any and all processes and therefore merely reveals the results of evolutionary history that have a fidelity in their information content. This means that they are equally applicable to any and all kinds of organisms, making obvious comparisons of numbers of species across taxa or geographical space or ecosystems possible.

History, Philosophy and Sociology of Taxonomy

It is important that professional philosophers, historians and social scientists be engaged in a rigorous analysis of taxonomy, particularly post-New Synthesis taxonomy, in order to understand and avoid a repetition of why support could be denied to the most fundamental and essential of all biodiversity sciences. I have repeatedly argued that the plight of contemporary taxonomy derives from a caustic political correctness founded in ignorance of the epistemic basis of taxonomy, unwarranted arrogance regarding experimentalism and the *tecnologie du jour* traceable to the New Systematics (see Wheeler, 1995a, 2004).

Engage Museums in Global Inventory

Collections are the most important product of taxonomy. If we could do only one thing, it would be to make museums a comprehensive reflection of species diversity on Earth as a baseline for future studies of evolution, phylogeny, species, morphology, ecosystem composition and biogeography. In order to expand and develop collections efficiently, however, it is necessary that taxonomic knowledge proceed in tandem with collection growth. Otherwise, it is impossible to know how comprehensive our collections are or to set priorities to assure the optimal species representation. This means that revisions must be done as the logical and efficient means to test and discover species. And revisions are accelerated through team work and improved and more efficient tools that allow access to specimens, characters, literature, data and expertise.

Democratization of Taxonomy

In addition to making taxonomy far more efficient at what it does, without the compromises in excellence inherent in alternatives like DNA taxonomy (see Lipscomb et al., 2003; Wheeler, 2005), taxonomy is democratized. Where in the past synoptic collections and comprehensive libraries were the province of a few privileged scientists in European and North American institutions, suddenly all and equal research resources exist for taxonomists, students and amateurs in remote places, small institutions and developing nations, where the bulk of species exploration is yet to be done.

From a taxon knowledge base, the user can not only define publications on demand but also produce cladograms on demand to his or her specifications. My first concern is that taxonomists have everything they need to work to the high standards that have been hard won over centuries while at the same time working smarter and faster. Unless taxonomists have the infrastructure they need, their many user communities will be disserved.

Users who are given access to accurate species identifications and all available information and interpretations about their characters, relationships, and distributions will quickly come to value taxonomy more. It is unreasonable to expect users to respect or value that which fails to meet their needs.

Museum Leadership

The New Taxonomy also needs a new generation of museum directors who have no doubt in their mind why their institutions exist or why we have and continue to build collections. Because specimens are so central to taxonomy, those institutions that are the caretakers of such collections have a special responsibility in the revival of taxonomy. Unless those leaders believe and can explain to politicians and the public that taxonomy matters, they will continue to be underfunded and will continue to fail to meet the demanding requirements of taxonomy itself. Unless revisions are done on an ongoing basis, the information content of such collections rapidly erodes and becomes unreliable to its users. Without information content, then the reason for society to maintain such vast collections evaporates. When taxonomy is ignored by museum directors, their *raison d'etre* is undermined and they might as well keep a small percentage of specimens of historical or curiosity interest and give up on the bulk of the collections. Those collections represent most of what we know or will ever come to know of biological diversity. Most museum directors appreciate their mandate to care for specimens' physical well-being, but many seem curiously ignorant of the equal necessity to care for their information content by assuring that taxonomy is a vibrant and healthy science.

CONCLUSIONS

The time has come to stop whinging about the biodiversity crisis and start to do something about it. Very little has changed since E.O. Wilson (1985) pointed out how little we know of Earth's species. His contraction biodiversity became a household word and the basis of international agreements while at the same time the science necessary for improving the situation continued to be ignored. A first and necessary step towards conserving, managing, using and enjoying biodiversity is to know it–not as reflected wavelengths of light detected by orbiting satellites, romantic remnants of beautiful places with unnumbered and unnamed occupants or faceless providers of essential ecological services, but rather as individually known and understood species. To know species we need to study them in detail, to enumerate their unique combinations of characters, to analyse their cladistic relationships and to describe, name and classify them.

The starry-eyed suggestion that some handheld DNA analysis tool will eliminate the burden of educating taxonomists and end the nuisance of needing to learn to recognize species or learn names for their characters (Hebert et al., 2003) is a dismal, bleak and defeated future. It is not coincidental that those suggesting such a technologically driven but intellectually vacuous future for taxonomy are non-taxonomists. To them, taxonomy is an impediment to their own needs and ambitions. They do not even pause to educate themselves about the intellectual riches of taxonomy or the philosophical rigor of taxonomy as an independent science. All that aside, it remains that the best taxonomic information comes from inspired and rigorous taxonomy. Like most cheap alternatives to excellence, the users of such arbitrary DNA distance alternatives to credible taxonomy will get what they pay for, but it may be too late to replace it with good taxonomy when many species and their characters have become extinct.

The time has come for taxonomists to pull together and demand the support that they need for research, education, collections and cyberinfrastructure in order to meet the biodiversity challenge.

The time has come to mount a taxonomic campaign of shock and awe. If the community bands together behind LINNE and similar taxonomic modernization activities to unveil a New Taxonomy, it can produce so much knowledge of the species of Earth and their characters, relationships and distributions that it will literally overwhelm any criticisms of the need to support museums, to support taxonomists and to avoid cheap substitutes for good science.

The public will be shocked to learn that the science community has suppressed the exploration of our world's species for so many decades. As taxonomists explore the biosphere the world will be awed at the diversity and beauty of Earth's species and the utility of knowledge about them. Trillions of dollars are associated with our use of 'other' species in agriculture, forestry, animal husbandry, plant breeding and trade in natural products including pharmaceuticals. That massive economy, however, is based on knowledge of perhaps no more than 10 per cent of living species. Imagine what clever humans could do with complete knowledge of Earth's species. Experimental biologists will find superior model organisms. Agriculturalists will have new and genetically vastly improved plants. Physicians will have whole new families of drugs. Homeland security officers will know when species cross our borders that do not belong here. Biologists will see and interpret phenomena in the context of evolutionary relationships. Future generations will marvel at the diversity of life on Earth in the early twenty-first century as they stand open jawed in vast natural history museums that are monuments to the jewels of billions of years of Earth's history. Every child will be able to venture into nature and identify and learn about every living marvel that he or she encounters. Amateurs and ecotourists will revel in their knowledge of species. Humanity, having seen the face and learned the names of their fellow species on this planet, will come to respect and value them precisely because they are accessible and knowable to them.

We need to set immediate and ambitious goals that will generate so much discovery and knowledge so rapidly that any arguments to conserve ecosystems now and study species later evaporate in the face of monumental understanding. That argument is merely an attempt to drain funding from the hard work of taxonomy into other areas of science. The reality is that we cannot save that which we do not know or know that we have even if we do. Arguments for ignorance collapse under the weight of knowledge.

We are the last generation with the opportunity to fully explore the diversity of life on this little-known planet. To do so we must invest heavily in taxonomy, taxonomists, collections and research infrastructure. We must retain the best of traditional taxonomic theory and practice and fuse it with appropriate technologies (and uses of technologies) from DNA sequencing to cyberinfrastructure. We must complete Hennig's taxonomic revolution and refocus our efforts on constructing and continually testing his general reference system. We must reorganize ourselves to elevate taxonomy to its proper status as a planetary-scale and independent science. Specialists must organize international knowledge communities that build and maintain comprehensive taxon knowledge banks from which all useful data, information and knowledge flows and into which all discoveries, revisions and advances are deposited. These communities must reconceive taxonomic revisions and monographs in the fast-paced realm of cybertaxonomy. Computer scientists and engineers must partner with taxonomists and museum specialists to identify and overcome each and every barrier to rapid advances without compromise to theory or quality of work. And taxonomists must get on with the urgent need to explore the species of Earth. Now is the time.

It is my hope that this volume marks a change in the history of taxonomy–a deliberate return to the bold visions of Linnaeus, Darwin and Hennig and a deliberate reversal of the ill effects of the New Systematics. Taxonomy must reassert itself as an independent science. From this science come the essential taxonomic information and services that subtend all credible basic and applied biology.

If we work together; if we work smart; if we speak with one voice; if we have clarity of purpose; if we are not distracted by those who would sacrifice quality for short-term gains; and if we undertake taxonomic exploration and research on a massive scale we can overcome prejudice and ignorance to succeed in our noble and necessary mission through a campaign of taxonomic shock and awe. No science has more fundamental and important discoveries within its grasp. No science has a more urgent or fundamental mission. No science has greater relevancy to human welfare and understanding. Although taxonomy is intellectually primarily an evolutionary science discovering species, clades and histories of character transformations, no science has more to contribute to environmental knowledge or conservation. Let us hope that 70 years hence biologists look back upon us as a generation with the vision to see the mistakes of its past, to avoid the mistakes of its present and to step up to the challenge to contribute to human knowledge that which we alone can do as a result of the biodiversity crisis: a comprehensive inventory of our world's species. Defeat is not an acceptable option and we have labored in the shadow of the New Systematics far too long. Shock and awe … make it so.

REFERENCES

Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Mesina, P., Ostriker, J.P. and Wright, M.H. (2003) Revolutionizing science and engineering through cyberinfrastructure. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Arlington, VA, US National Science Foundation, 84 pp.

Carpenter, J. (2003) A critique of pure folly. *Botanical Review,* 69: 79–92.

Cracraft, J. (2002) The seven great questions of systematic biology: An essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanical Garden,* 89: 127–144.

Desalle, R. (2006) Species discovery versus species identification in DNA barcoding efforts: Response to Rubinoff. *Conservation Biology,* 20: 1545–1547.

Dominguez, E. and Wheeler, Q.D. (1997) Taxonomic stability is ignorance. *Cladistics,* 13: 367–372.

Ebach, M.C. and Holdrege, C. (2005) DNA barcoding is no substitute for taxonomy. *Nature,* 434: 697.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and de Waard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B,* 270: 313–322.

Hennig, W. (1966) *Phylogenetic Systematics,* University of Illinois Press, Urbana, 263 pp.

Keller, R.A., Boyd, R.N. and Wheeler, Q.D. (2003) The illogical basis of phylogenetic nomenclature. *Botanical Review,* 69: 83–110.

Lipscomb, D.N., Platnick, N.I. and Wheeler, Q.D. (2003) The intellectual content of taxonomy: A comment on DNA taxonomy. *Trends in Ecology and Systematics,* 18: 65–66.

Little, D.P. and Stevenson, D.W. (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: Examples from gymnosperms. *Cladistics,* 23: 1–21.

MacLeod, N., ed (2007) *Automated Taxon Identification in Systematics: Theory, Approaches and Applications,* CRC Press, Boca Raton, FL, 339 pp.

Meier, R. and Dikow, T. (2004) The significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable and complete specimen data. *Conservation Biology,* 18: 478–488.

Meier, R., Kwong, S., Vaidya, G. and Ng, P.K.L. (2006) DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology,* 55: 715–728.

Meyer, C.P. and Paulay, G. (2005) DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology,* 3: 2229–2238.

Miller, K.B., Alarie, Y., Wolfe, G.W. and Whiting, M.F. (2005) Association of insect life stages using DNA sequences: The larvae of *Philodytes umbrinus* (Motschulsky) (Coleoptera: Dytiscidae). *Systematic Entomology,* 30: 499–509.

Nelson, G. and Platnick, N.I. (2001) *Systematics and Biogeography: Cladistics and Vicariance,* Columbia University Press, New York, 567 pp.

Nixon, K.C. and Wheeler, Q.D. (1992) Extinction and the origin of species. In *Extinction and Phylogeny* (eds M.J. Novacek and Q.D. Wheeler), Columbia University Press, New York, pp. 119–143.

Nixon, K.C., Carpenter, J.M. and Stevenson, D.W. (2003) The PhyloCode is fatally flawed, and the 'Linnaean' system can easily be fixed. *Botanical Review,* 69: 111–120.

Page, L.M., Bart, H.L., Jr., Beaman, R., Bohs, L., Deck, L.T., Funk, V.A., Lipscomb, D., Mares, M.A., Prather, L.A., Stevenson, J., Wheeler, Q.D., Woolley, J.B. and Stevenson, D.W. (2005) *LINNE: Legacy Infrastructure Network for Natural Environments,* Illinois Natural History Survey, Champaign, Illinois, 16 pp.

Polaszek, A., Agosti, D., Alonso-Zarazaga, M., Beccaloni, G., Bjorn, P., Bouchet, P., Brothers, D.J., Cranbrook, G., Evenhuis, N., Godfray, C.J., Johnson, N.F., Krell, F.T., Lipscomb, D., Lyal, C.H.C., Mace, G.M., Mawatari, S., Miller, S.E., Minelli, A., Morris, S., Ng, P.K.L., Patterson, D.J., Pyle, R.L., Robinson, N., Rogo, L., Taverne, J., Thompson, F.C., Tol, J., Wheeler, Q.D. and Wilson, E.O. (2005) A universal register for animal names. *Nature,* 437: 477.

Prendini, L. (2005) Comments on 'Identifying spiders through DNA barcodes'. *Canadian Journal of Zoology,* 83: 498–504.

Schuh, R.T. (2000) *Biological Systematics,* Cornell University Press, Ithaca, New York, 256 pp.

Ullman, H.K. and Wade, J.P. (1996) *Shock and Awe: Achieving Rapid Dominance,* National Defense University, Washington, D.C., 142 pp.

Wheeler, Q.D. (1995) The 'Old Systematics': Classification and phylogeny. In *Biology, Phylogeny, and Classification of Coleoptera. Papers Celebrating the 80th Birthday of Roy A. Crowson* (eds J. Pakaluk and S.A. Slipinski), Muzeum i Insytut Zoologii PAN, Warszawa, pp. 31-62.

Wheeler, Q.D. (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society of London, Series B,* 359: 571–583.

Wheeler, Q.D. (2005) Losing the plot: DNA barcodes and taxonomy. *Cladistics,* 21: 405–407.

Wheeler, Q.D. (2007) Digital innovation and taxonomy's finest hour. In *Automated Taxon Identification in Systematics: Theory, Approaches and Applications* (ed N. MacLeod), CRC Press, Boca Raton, FL, pp. 9–23.

Wheeler, Q.D. and Platnick, N.I. (2000) The phylogenetic species concept. In *Species Concepts and Phylogenetic Theory: A Debate* (ed Q.D. Wheeler), Columbia University Press, New York, pp. 55–69.

Wheeler, Q.D., Raven, P.H. and Wilson, E.O. (2004) Taxonomy: Impediment or expedient? *Science,* 303: 285.

Will, K.P., Mishler, B.D. and Wheeler, Q.D. (2005) The perils of DNA bar-coding and the need for integrative taxonomy. *Systematic Biology,* 54: 844–851.

Wilson, E.O. (1985) The biological diversity crisis: A challenge to science. *Issues in Science and Technology,* 2: 20–29.

Index

Systematics Association Publications

1. Bibliography of Key Works for the Identification of the British Fauna and Flora, 3rd edition (1967)†
   Edited by G.J. Kerrich, R.D. Meikie and N. Tebble
2. Function and Taxonomic Importance (1959)†
   Edited by A.J. Cain
3. The Species Concept in Palaeontology (1956)†
   Edited by P.C. Sylvester-Bradley
4. Taxonomy and Geography (1962)†
   Edited by D. Nichols
5. Speciation in the Sea (1963)†
   Edited by J.P. Harding and N. Tebble
6. Phenetic and Phylogenetic Classification (1964)†
   Edited by V.H. Heywood and J. McNeill
7. Aspects of Tethyan Biogeography (1967)†
   Edited by C.G. Adams and D.V. Ager
8. The Soil Ecosystem (1969)†
   Edited by H. Sheals
9. Organisms and Continents through Time (1973)†
   Edited by N.F. Hughes
10. Cladistics: A Practical Course in Systematics (1992)*
    P.L. Forey, C.J. Humphries, I.J. Kitching, R.W. Scotland, D.J. Siebert and D.M. Williams
11. Cladistics: The Theory and Practice of Parsimony Analysis (2nd edition) (1998)*
    I.J. Kitching, P.L. Forey, C.J. Humphries and D.M. Williams

* Published by Oxford University Press for the Systematics Association

† Published by the Association (out of print)

SYSTEMATICS ASSOCIATION SPECIAL VOLUMES

1. The New Systematics (1940)
   Edited by J.S. Huxley (reprinted 1971)
2. Chemotaxonomy and Serotaxonomy (1968)*
   Edited by J.C. Hawkes
3. Data Processing in Biology and Geology (1971)*
   Edited by J.L. Cutbill

4. Scanning Electron Microscopy (1971)*
Edited by V.H. Heywood

5. Taxonomy and Ecology (1973)*
Edited by V.H. Heywood

6. The Changing Flora and Fauna of Britain (1974)*
Edited by D.L. Hawksworth

7. Biological Identification with Computers (1975)*
Edited by R.J. Pankhurst

8. Lichenology: Progress and Problems (1976)*
Edited by D.H. Brown, D.L. Hawksworth and R.H. Bailey

9. Key Works to the Fauna and Flora of the British Isles and Northwestern Europe, 4th edition (1978)*
Edited by G.J. Kerrich, D.L. Hawksworth and R.W. Sims

10. Modern Approaches to the Taxonomy of Red and Brown Algae (1978)
Edited by D.E.G. Irvine and J.H. Price

11. Biology and Systematics of Colonial Organisms (1979)*
Edited by C. Larwood and B.R. Rosen

12. The Origin of Major Invertebrate Groups (1979)*
Edited by M.R. House

13. Advances in Bryozoology (1979)*
Edited by G.P. Larwood and M.B. Abbott

14. Bryophyte Systematics (1979)*
Edited by G.C.S. Clarke and J.G. Duckett

15. The Terrestrial Environment and the Origin of Land Vertebrates (1980)
Edited by A.L. Pachen

16 Chemosystematics: Principles and Practice (1980)*
Edited by F.A. Bisby, J.G. Vaughan and C.A. Wright

17. The Shore Environment: Methods and Ecosystems (2 volumes) (1980)*
Edited by J.H. Price, D.E.C. Irvine and W.F. Farnham

18. The Ammonoidea (1981)*
Edited by M.R. House and J.R. Senior

19. Biosystematics of Social Insects (1981)*
Edited by P.E. House and J.-L. Clement

20. Genome Evolution (1982)*
Edited by G.A. Dover and R.B. Flavell

21. Problems of Phylogenetic Reconstruction (1982)
Edited by K.A. Joysey and A.E. Friday

22. Concepts in Nematode Systematics (1983)*
Edited by A.R. Stone, H.M. Platt and L.F. Khalil

23. Evolution, Time and Space: The Emergence of the Biosphere (1983)*
Edited by R.W. Sims, J.H. Price and P.E.S. Whalley

24. Protein Polymorphism: Adaptive and Taxonomic Significance (1983)*
    Edited by G.S. Oxford and D. Rollinson

25. Current Concepts in Plant Taxonomy (1983)*
    Edited by V.H. Heywood and D.M. Moore

26. Databases in Systematics (1984)*
    Edited by R. Allkin and F.A. Bisby

27. Systematics of the Green Algae (1984)*
    Edited by D.E.G. Irvine and D.M. John

28. The Origins and Relationships of Lower Invertebrates (1985)‡
    Edited by S. Conway Morris, J.D. George, R. Gibson and H.M. Platt

29. Infraspecific Classification of Wild and Cultivated Plants (1986)‡
    Edited by B.T. Styles

30. Biomineralization in Lower Plants and Animals (1986)‡
    Edited by B.S.C. Leadbeater and R. Riding

31. Systematic and Taxonomic Approaches in Palaeobotany (1986)‡
    Edited by R.A. Spicer and B.A. Thomas

32. Coevolution and Systematics (1986)‡
    Edited by A.R. Stone and D.L. Hawksworth

33. Key Works to the Fauna and Flora of the British Isles and Northwestern Europe, 5th edition (1988)‡
    Edited by R.W. Sims, P. Freeman and D.L. Hawksworth

34. Extinction and Survival in the Fossil Record (1988)‡
    Edited by G.P. Larwood

35. The Phylogeny and Classification of the Tetrapods (2 volumes) (1988)‡
    Edited by M.J. Benton

36. Prospects in Systematics (1988)‡
    Edited by J.L. Hawksworth

37. Biosystematics of Haematophagous Insects (1988)‡
    Edited by M.W. Service

38. The Chromophyte Algae: Problems and Perspective (1989)‡
    Edited by J.C. Green, B.S.C. Leadbeater and W.L. Diver

39. Electrophoretic Studies on Agricultural Pests (1989)‡
    Edited by H.D. Loxdale and J. den Hollander

40. Evolution, Systematics, and Fossil History of the Hamamelidae (2 volumes) (1989)‡
    Edited by P.R. Crane and S. Blackmore

41. Scanning Electron Microscopy in Taxonomy and Functional Morphology (1990)‡
    Edited by D. Claugher

42. Major Evolutionary Radiations (1990)‡
    Edited by P.D. Taylor and G.P. Larwood

43. Tropical Lichens: Their Systematics, Conservation and Ecology (1991)[‡]
    Edited by G.J. Galloway

44. Pollen and Spores: Patterns and Diversification (1991)[‡]
    Edited by S. Blackmore and S.H. Barnes

45. The Biology of Free-Living Heterotrophic Flagellates (1991)[‡]
    Edited by D.J. Patterson and J. Larsen

46. Plant–Animal Interactions in the Marine Benthos (1992)[‡]
    Edited by D.M. John, S.J. Hawkins and J.H. Price

47. The Ammonoidea: Environment, Ecology and Evolutionary Change (1993)[‡]
    Edited by M.R. House

48. Designs for a Global Plant Species Information System (1993)[‡]
    Edited by F.A. Bisby, G.F. Russell and R.J. Pankhurst

49. Plant Galls: Organisms, Interactions, Populations (1994)[‡]
    Edited by M.A.J. Williams

50. Systematics and Conservation Evaluation (1994)[‡]
    Edited by P.L. Forey, C.J. Humphries and R.I. Vane-Wright

51. The Haptophyte Algae (1994)[‡]
    Edited by J.C. Green and B.S.C. Leadbeater

52. Models in Phylogeny Reconstruction (1994)[‡]
    Edited by R. Scotland, D.I. Siebert and D.M. Williams

53. The Ecology of Agricultural Pests: Biochemical Approaches (1996)**
    Edited by W.O.C. Symondson and J.E. Liddell

54. Species: the Units of Diversity (1997)**
    Edited by M.F. Claridge, H.A. Dawah and M.R. Wilson

55. Arthropod Relationships (1998)**
    Edited by R.A. Fortey and R.H. Thomas

56. Evolutionary Relationships among Protozoa (1998)**
    Edited by G.H. Coombs, K. Vickerman, M.A. Sleigh and A. Warren

57. Molecular Systematics and Plant Evolution (1999)
    Edited by P.M. Hollingsworth, R.M. Bateman and R.J. Gornall

58. Homology and Systematics (2000)
    Edited by R. Scotland and R.T. Pennington

59. The Flagellates: Unity, Diversity and Evolution (2000)
    Edited by B.S.C. Leadbeater and J.C. Green

60. Interrelationships of the Platyhelminthes (2001)
    Edited by D.T.J. Littlewood and R.A. Bray

61. Major Events in Early Vertebrate Evolution (2001)
    Edited by P.E. Ahlberg

62. The Changing Wildlife of Great Britain and Ireland (2001)
    Edited by D.L. Hawksworth

63. Brachiopods Past and Present (2001)
    Edited by H. Brunton, L.R.M. Cocks and S.L. Long
64. Morphology, Shape and Phylogeny (2002)
    Edited by N. MacLeod and P.L. Forey
65. Developmental Genetics and Plant Evolution (2002)
    Edited by Q.C.B. Cronk, R.M. Bateman and J.A. Hawkins
66. Telling the Evolutionary Time: Molecular Clocks and the Fossil Record (2003)
    Edited by P.C.J. Donoghue and M.P. Smith
67. Milestones in Systematics (2004)
    Edited by D.M. Williams and P.L. Forey
68. Organelles, Genomes and Eukaryote Phylogeny (2004)
    Edited by R.P. Hirt and D.S. Horner
69. Neotropical Savannas and Seasonally Dry Forests: Plant Diversity, Biogeography and
    Conservation (2006)
    Edited by R.T. Pennington, G.P. Lewis and J.A. Rattan
70. Biogeography in a Changing World (2006)
    Edited by M.C. Ebach and R.S. Tangney
71. Pleurocarpous Mosses: Systematics & Evolution (2006)
    Edited by A.E. Newton and R.S. Tangney
72. Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa (2006)
    Edited by T.R. Hodkinson and J.A.N. Parnell
73. Biodiversity Databases: Techniques, Politics, and Applications (2007)
    Edited by G.B. Curry and C.J. Humphries
74. Automated Taxon Identification in Systematics: Theory, Approaches and Applications (2007)
    Edited by N. MacLeod
75. Unravelling the algae: the past, present, and future of algal systematics (2008)
    Edited by Juliet Brodie and Jane Lewis

[*] Published by Academic Press for the Systematics Association

[†] Published by the Palaeontological Association in conjunction with Systematics Association

[‡] Published by the Oxford University Press for the Systematics Association

[**] Published by Chapman & Hall for the Systematics Association

The New Taxonomy

It's not just curiosity about the world that fuels the taxonomic fire. The most fundamental of all biological sciences, taxonomy underpins any long term strategies for reconstructing the great tree of life or salvaging as much biodiversity as possible. Yet we are still unable to say with any certainty how many species are living on the Earth. **The New Taxonomy** describes how a confluence of theory, cyberinfrastructure and international teamwork can meet this unprecedented research challenge and marks an emerging field, cybertaxonomy.

The book examines the efforts of several international groups to catalog the world's biodiversity and make it accessible. An answer to Julian Huxley's *The New Systematics*, the book signals the beginning of an upward trajectory of taxonomy to meet the unprecedented challenges of the biodiversity crisis. Contemporary taxonomists reclaim the unique mission, goals and importance of taxonomy as an independent science. They cover technologies such as DNA evidence and its applications, computer-assisted species identification, digital morphology and E-typification.

Not much has changed since E.O. Wilson pointed out how little we know of the Earth's species in 1985. This book offers a vision and a strategy for changing all that. The first current, unapologetic look at morphology and descriptive taxonomy that points out their incredible importance to science and society, this book frames one of the most constructive responses to the biodiversity crisis.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

9088

ISBN: 978-0-8493-9088-3
90000
9 780849 390883
www.crcpress.com