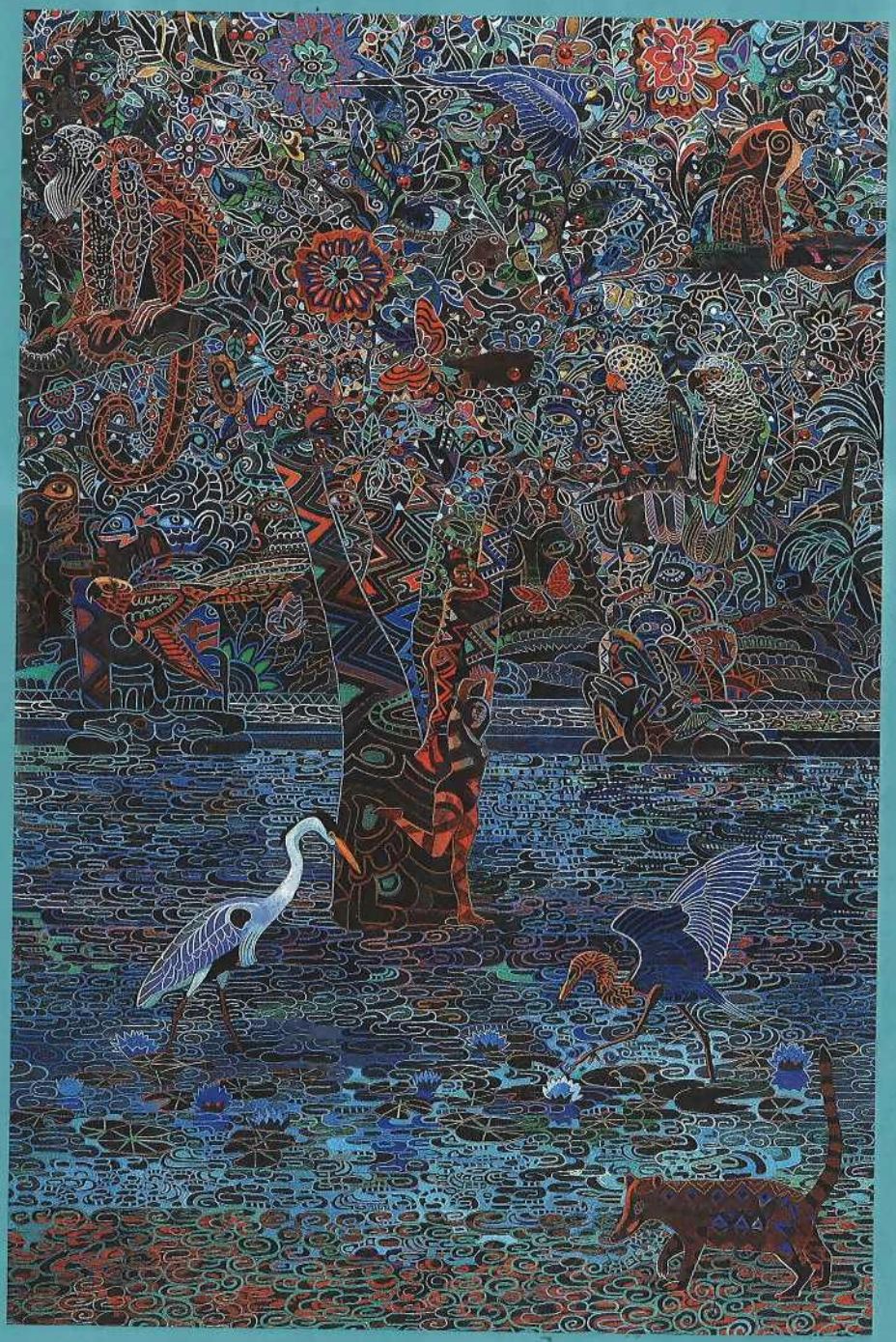
BIODIVERSITY II

Understanding and Protecting Our Biological Resources



Marjorie L. Reaka-Kudla, Don E. Wilson,
and Edward O. Wilson, editors

CHAPTER
13

Names: The Keys to Biodiversity

F. CHRISTIAN THOMPSON
Research Entomologist, Systematic Entomology Laboratory, Plant Sciences Institute, Beltsville Agricultural Research Center, United States Department of Agriculture, Washington, D.C.

Besides biodiversity, the one thing that all the chapters included in this volume have in common is scientific names. These names form the essential language, the means we use to communicate about biodiversity. To avoid a Tower of Babel, a common system of nomenclature is required: a system that is effective and efficient (and at minimal cost). Presented below are the essential aspects of this language for biodiversity and a discussion of where we are in respect to their implementation.

The long-term conservation of biodiversity can be achieved only through the approach used by the Instituto Nacional de Biodiversidad (INBio)—"save it," "characterize it," and "sustainably use it" (Janzen, Chapter 27 of this volume). Characterization requires that we have a language with which to communicate about biodiversity: a way of describing it, so that we all know what we are talking about and that we are talking about the same things. How do we characterize biodiversity? The first step is to name its components. Biodiversity is divisible into three levels: ecological, taxonomic, and genetic. Of these levels, taxonomic diversity is critical because taxa are the units that contain genetic diversity and are the units that make up ecological diversity. Since taxa are the core of biodiversity, names for taxa are the most critical component of any language of biodiversity.

A UNIVERSAL LANGUAGE OF BIODIVERSITY

What are names? Names are tags. Tags are words, short sequences of symbols that are used in place of something complex which would require many

more words to describe. Hence, tags save time and space. Instead of a long description, we use a short tag. A scientific name differs from a common name in that the scientific name is a unique tag. In other languages, there may be multiple tags for the same thing. Imagine the various words in English that are used to describe *Homo sapiens*. In computer (database) jargon, data elements that are used to index information are termed keys, and keys that are unique are called primary keys. Scientific names are primary keys. The word "key" has another meaning in English, which is "something that unlocks something." Scientific names are those critical keys that unlock biosystematic information, all that we know about living organisms. Scientific names are tags that replace descriptions of objects or, more precisely, concepts based on objects (specimens). Scientific names are unique, there being only one scientific name for a particular concept, and each concept has only one scientific name.

Scientific names are more than just primary keys to information. They represent hypotheses. To systematists, this is a trivial characteristic that sometimes is forgotten and thereby becomes a source of confusion later. To most users, this is an unknown characteristic that prevents them from obtaining the full value from scientific names. If a scientific name were only a unique key used for storing and retrieving information, it would be just like a social security number. *Homo sapiens* is a unique key used to store and retrieve information about man, but that key also places the information about man into a hierarchical classification. Hierarchical classifications allow for the storage, at each node of the hierarchy, of information common to the subordinate nodes. Hence, redundant data, which would be spread throughout a nonhierarchical system, are eliminated.

Biological classifications, however, do more than just hierarchically store information. Given that one accepts a single common (unique) history for life and that our biological classifications reflect this common history in their hierarchical arrangement, then biological classifications allow for prediction: they allow us to predict that some information stored at a lower hierarchical node may belong to a higher node, that is, is common to all members of the more inclusive group. These predictions take the form of the following: if some members of a group share a characteristic that is unknown for other members of the same group, then that characteristic is likely to be common to all members of the group.

So scientific names are tags, unique keys, hierarchical nodes, and phylogenetic hypotheses. Thus, systematists pack a lot of information into their names and users can get a lot from them.

Scientific names are hypotheses, not proven facts. Systematists may and frequently do disagree about hypotheses. Hypotheses, which in systematics range from what is a character to what is the classification that best reflects the history of life, are always prone to falsification, and, hence, change. Disagreements about classification can arise from differences in paradigm or information.

Systematists use different approaches to construct classifications, such as cladistic versus phylogenetic versus phenetic methods. Given the same set of data that underlies a given hierarchy, cladists can derive classifications that are different from those derived by the phenetists (Figure 13-1). Even among cladists, there can be differences as to the rank (genus, family, order, etc.) and thereby the hierarchical groups used. These are disagreements based on paradigm. There can be disagreement about the hypotheses that underlie the information used to construct the classifications, such as what are the characters. Disagreement can arise among systematists because they use different information. While disagreements will affect the ability to predict, they need not affect the ability to retrieve information.

The desirable attribute that must be preserved to ensure complete access to information across multiple classifications is uniqueness. Our scientific nomenclature must guarantee that any scientific name that is used in any classification is unique among all classifications. This can be assured by having two primary keys. Unfortunately, having two keys increases the overhead of our information systems. Most systematists and *all* users want to avoid this problem by

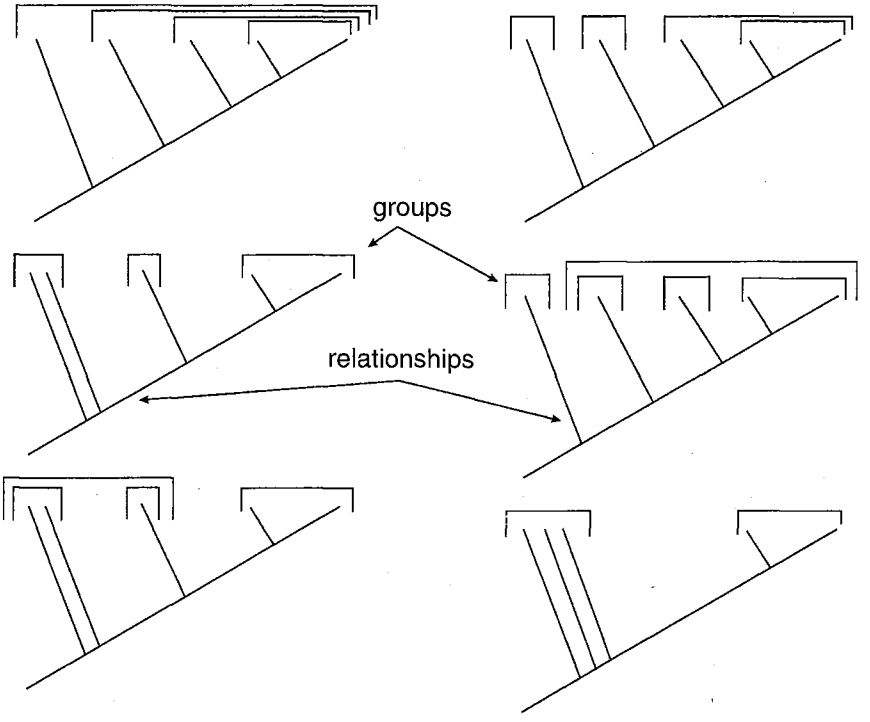


FIGURE 13-1 Multiple classifications for identical cladistic hypotheses.

mandating that there be only *one* classification. Although in theory there is only one correct classification, as there was only one history of life, in reality there have been multiple classifications in the past, there may be multiple classifications in use today, and there will be multiple classifications in the future. That is the price of scientific progress, of the increase in our knowledge of the world. If information is to be retrieved across time, if we want to extract information stored under obsolete classifications, and if we want to avoid dictating "the correct" classification, then we need a nomenclatural system that supports two unique keys.

The two keys for our language of biodiversity are the *valid* name and the *original* name. The valid name is the correct name for a concept within a classification; the original name is the valid name in the classification in which it was proposed. Valid names may be different among classifications, but the original name is invariant across classifications (Table 13-1). Valid names are the best names to use because they provide the full value of scientific names. These are the names that provide a basis for prediction. The original name is useful only for retrieval of information across multiple classifications. Although valid and original names may be and frequently are the same, users must know the differences between them. Specifically, they need to know that a valid name is a powerful tool for inference, that a valid name provides for prediction of unknown attributes of the organism that bears the name. But they must understand that there may be multiple valid names in the literature or in use, and that valid names represent hypotheses that may change as our knowledge is tested and improved. Most importantly, if there are multiple valid names in use, then the users must be aware that there are conflicting scientific hypotheses being advocated and that they must select the name that best serves their purpose. If users do not want to decide, do not want to use classifications to organize and synthesize their information, then they may use the original name to index their information, being assured that it will always be a unique key.

There are other problems today with our classifications: synonymy (having two names for the same concept) and homonymy (having the same name for

TABLE 13-1 Multiple Classifications and Primary Keys to Information

Year	Valid Name	Original Name	Authority
1776	*Musca balteata*	*Musca balteata*	De Geer
1822	*Syrphus balteatus*	*Musca balteata*	Meigen
1843	*Scaeva balteatus*	*Musca balteata*	Zetterstedt
1917	*Episyrphus balteatus*	*Musca balteata*	Matsumura
1930	*Epistrophe balteata*	*Musca balteata*	Sack
1950	*Stenosyrphus balteatus*	*Musca balteata*	Fluke
Today	*Episyrphus balteatus*	*Musca balteata*	Vockeroth

different concepts). These problems are, however, largely due to ignorance. If we knew all names and their types and could agree on what are species, then by applying the rules of nomenclature we immediately could eliminate all problems of synonymy and homonymy. Homonymy is eliminated by the rule of uniqueness. Synonymy is addressed by the rules of typification, which tie a physical instance of a concept to a name. Synonymy is resolved by logic of circumscription and the convention of priority (or usage). The name of a concept is the name affixed to one and only one of the types that falls within its circumscription. The name used is determined by which name is the oldest (priority) or most widely used (usage). The specific rules for resolving homonymy and synonymy, as well as for the proper formation and documentation of names, are our Codes of Nomenclature (International Botanical Congress, 1994; International Commission on Zoological Nomenclature, 1985; Sneath, 1992). These rules, however, do not address the problem of multiple classifications or ignorance of the universe of applicable names and their typification.

There is one final problem, the species problem. This is the problem of the basic unit of information or data. The basic unit of information for nomenclature is the species (or more precisely the species-group, which includes the category of subspecies). The problem is that the species is not a single element of datum, but consists of information—data derived from specimens that have been identified as belonging to that species. Mistakes can be made during this identification, which is another hypothesis. Information ultimately is not derived from species, but from specimens. Information management in biodiversity really begins with data management of specimens. The problems of specimen-based data management are not intractable, but are readily addressed by the use of bar-codes, another form of unique keys.

The species problem is also one of circumscription, the definition of the limits of a taxon. A group with the same name and type may be more or less inclusive, depending on the characters used to define its limits. Zoologists differ from botanists in not considering circumscription to be a problem because, minimally, all identically named taxa have at least some characteristics in common. The problem of how much is held in common, therefore, is best resolved by enumeration of the included taxa or specimens. The history of circumscription can be tracked by use of an additional key that uniquely identifies the person who defined the limits and the date of that action. It is sufficient for our purposes to know that data based on specimens always will be summarized into information units based on species and that all such information should be based on specimens.

THE REAL WORLD: DIVERSITY AND DISPERSION

The All Taxa Biodiversity Inventories (ATBIs), taxasphere, national biological surveys, etc., are means of addressing the problem of characterization of

biodiversity, the most important step in conserving biodiversity. To characterize, a language is needed. To characterize biodiversity, all our available resources will be needed because the task is immense (World Conservation Monitoring Centre, 1992). The resources for characterizing biodiversity are diverse and dispersed. The only way the job is going to get done is by forming partnerships, by working together. That requires communications that can accommodate the diversity and dispersion.

If there were only one classification, if that classification were controlled by one person, and if all information were stored in one system at one place under that classification, then there would be one source for answers to questions about biodiversity. Unique and comprehensive information systems are powerful tools. In reality, however, things are different. Different classifications exist and resources are dispersed. To share resources, therefore, different classifications must be understood. To utilize distributed resources, a universal communication system that allows multiple classifications is needed. Scientific nomenclature provides the basis for this system: a set of unique keys to traverse the world of distributed databases, to find information anywhere on our future information superhighways. However, to make the system work, a universal data dictionary is required. Such a dictionary would allow users with any name to find the keys to unlock information about biodiversity. A universal data dictionary requires that the problems of synonymy and homonymy are solved and ensures that all classifications and names are accommodated.

PROGRESS AND PROMISE

The present language of biodiversity is binominal nomenclature that was introduced by Carolus Linnaeus, a Swedish professor of natural history. This system was the direct result of an earlier governmental biodiversity project. The Swedish Crown had some far-flung possessions and wanted to know what use could be made of them. They sent Linnaeus to investigate, to survey what today is called biodiversity, and to write a report characterizing what was found, with recommendations on how to use it. At the time, there was only a binary system of nomenclature: one word for the genus, with the species being described by a series of adjectives. Given the diversity Linnaeus found, he did not want to waste time repeating the long strings of adjectives that were required to characterize the biodiversity. Because the characterizations were in his flora of Sweden, he used a combination of the genus name and a single word (an epithet) for each species to form a unique key to those descriptions (see Stearn, 1957, for more details). The system was an immediate success. Linnaeus codified the system, built and maintained a universal information database for all names (his *Systema Naturae*, 1758), and trained a cadre of students to carry on his work. The students dispersed and converted others. But because no one could be *the* master except Linnaeus, they divided nature up. There were to be no more

Systema Naturae. At first, there were a series of *Systemae* for parts of nature, maintained by the students as the authorities for this kingdom or that area, but authorities in many countries quickly became involved. Two hundred years later, systematists cannot tell society how many organisms have been described or what are all their names.

What can systematists do? They can cooperate and, as a whole, recreate what Linnaeus started: a system of nature. They can agree on and follow a set of standards for nomenclature; ensure that those standards are adequate for our informational needs; eliminate the chaos of the past, first by gathering together the names that are today dispersed across a vast sea of literature, second by putting a limit on searching the past, and finally by accepting what is found after a reasonable search. They can ensure that the chaos will not return by requiring wide dissemination or registration of new names.

Where are systematists today in accomplishing these tasks? On standards, we are almost there. For bacteria, there is already a modern system of nomenclature (Ride, 1991; Sneath, 1986). For zoology and botany, we will have one shortly, and a start has been made on a universal code for all life (Hawksworth, 1994). Information about proposed changes in the botanical code are published regularly in the journal *Taxon*, and those for the zoological code in *Bulletin of Zoological Nomenclature*. A new draft version of the *International Code of Zoological Nomenclature* is now available for comment and can be adopted as soon as a year hence. With the acceptance of this new *Code*, the problems of name changes due to nomenclature will be eliminated. Some systematists are afraid that the new *Code* will be used to enforce the acceptance of one classification, rather than allow for a diversity of them. Such fears are unfounded because the new draft will preserve the "freedom of taxonomic thought and action." The draft includes new requirements stating that zoologists must document properly their classifications (implicit typification will be required for all names) and publish them where the whole community can evaluate them. This will be achieved by a registration system for new names. Bacteriology already requires this and botany has adopted it for the future. The new draft will be simpler to use, since the strict requirement of Latin grammar will be eliminated. Stability and universality will be enhanced by allowing zoologists to balance usage and priority more effectively in the determination of valid names. Finally, the new draft will provide the means to certify names and the associated nomenclatural data en masse, so as to free zoologists of the burdens of historical searches.

The last 250 years have left scientific names scattered across the most diverse array of media possible. No other science requires its practitioners to be responsible for such a mess. Scientists are expected to know the common and current hypotheses. They should not be required to know what was printed 200 years ago, distributed, and subsequently forgotten and largely lost. Systematists must deal with such ancient history, regardless of whether the concept had been published previously, was rediscovered subsequently, or was invalid.

That forgotten name need not even be in the domain of the systematist's expertise to cause problems.

The solution for systematics is simple: change the rules of nomenclature. This is being done. Then make a reasonable effort to gather together all the existing names and associated data and accept those names and data as correct. That will free systematists of the burden of history for nomenclature's sake.

The Systematic Entomology Laboratory of the U.S. Department of Agriculture has embarked on a voyage to do just this for insect names. We have proposed to the entomological community that together we develop a comprehensive database of names of terrestrial arthropods. We christened our ship BIOTA (Biosystematic Information on Terrestrial Arthropods). The most immediate and highest priority is to document all the names of terrestrial arthropods known to occur in North America. This represents the official adoption of the goal of the Entomological Collections Network, originally proposed by Miller (1992). We expect to reach this point within 2 years. We already have accumulated nearly 100,000 names and have commitments from cooperating specialists for another 40,000. This nomenclatorial database will include the essential keys, both the valid name and the original name for each species, that our specialists have recognized. The specialists will be identified in the data record, and minimal classificatory data, such as subfamily, family, order, and class, will be included. All synonyms, homonyms (all available names, sensu Zoological Code [International Commission on Zoological Nomenclature, 1985]) and common invalid combinations (names valid under other classifications) and misidentifications will be included. This information will provide the community with the necessary keys to the biodiversity of arthropods.

BIOTA also includes more comprehensive cataloging efforts, such as the Biosystematic Database of World Diptera. These efforts will include data on typification of names. The resultant catalogs will be submitted to the appropriate user-community for review and eventually to the International Commission of Zoological Nomenclature for certification. The Diptera database project has been endorsed by the International Congresses of Dipterology and is overseen by a committee of the Council for those Congresses (the council is a scientific member of the International Union of Biological Sciences). The family-group names, those names that may apply to the higher levels of classification, are nearing completion as the result of a 50-year effort by Curtis W. Sabrosky, one of our U.S.D.A. specialists. Some 4,296 names have been documented. Genus-group names have been entered and shortly will be distributed to specialists. Some 17,271 names were found, representing perhaps 8,000 valid genera. Names of species-groups are being entered now, with some 45,994 already in the database, perhaps 40% of the world total. All names for the flies of the Nearctic region have been entered and will be published shortly (28,890 names for 19,562 species and 2,356 genera).

Similar efforts are underway or already completed for other major groups of

organisms (Bisby, 1994). Given continued support, we are likely to see the problems of names solved by the year 2000. Nomenclature no longer will be an impediment to efforts to characterize our biodiversity. We then will have a set of keys to what we know of biodiversity and a language capable of effectively and efficiently incorporating new information and accommodating diverse keys to that information.

SERVICE TO SOCIETY

Names are the keys to biodiversity, but what does one do when one has no name, only a specimen? How does one discover the proper name for a specimen? If it is a bird, one can use a field guide, like Peterson's (1980), to identify it. If not, one asks an expert. If it is an insect, one will find those experts at the Systematic Entomology Laboratory (SEL), which identifies more insects than any other organization. But identification has costs: when experts identify specimens, they are not building classifications nor describing new biodiversity. Realizing the existing shortage of experts to classify specimens, SEL sought new approaches to relieve their experts of the burdens and distractions of identification.

The obvious answer was found within the question: if one does not need experts to identify birds, why does one need them to identify flies? Peterson (1980) proved that, if users were presented with the critical characters (his field marks) in a graphic way, then users readily could identify birds. Many of the field marks used in Peterson's field guides were known to Linnaeus 200 years ago, but Linnaean descriptions are difficult for users to understand. Compare, for example, the plate of common ducks in Peterson's field guide (Peterson, 1980:51) and the corresponding page from Linnaeus' *Systema Naturae* (Linnaeus, 1758: 126) (Figure 13-2). As good as Peterson field guides are, they, like the traditional identification key, are rather inflexible. To identify, for example, a pintail, one must first know that it is a duck, a freshwater dabbler, or one must thumb through the pages of the field guide until the appropriate picture is found. With a computerized identification system, the user can select the most obvious character, such as the long tail. Then the computer would list the dozen or so species that have long tails. The user could ask the computer what are the best characters to discriminate among these species, and the computer would respond with a ranked list of the most useful characters, some of which may be head color, body shape, habitat, and geography.

A computerized identification system that builds on and extends the visual approach of Peterson has been produced. The Fruit Fly Expert System allows users to identify some 200 fruit flies, including all the important species of pests (Thompson et al., 1993). The system uses data encoded in the standard DELTA (DEscriptive Language for TAxonomy; Pankhurst, 1991) format, so other data sets can be prepared easily. The system is extremely flexible. Many taxa can be

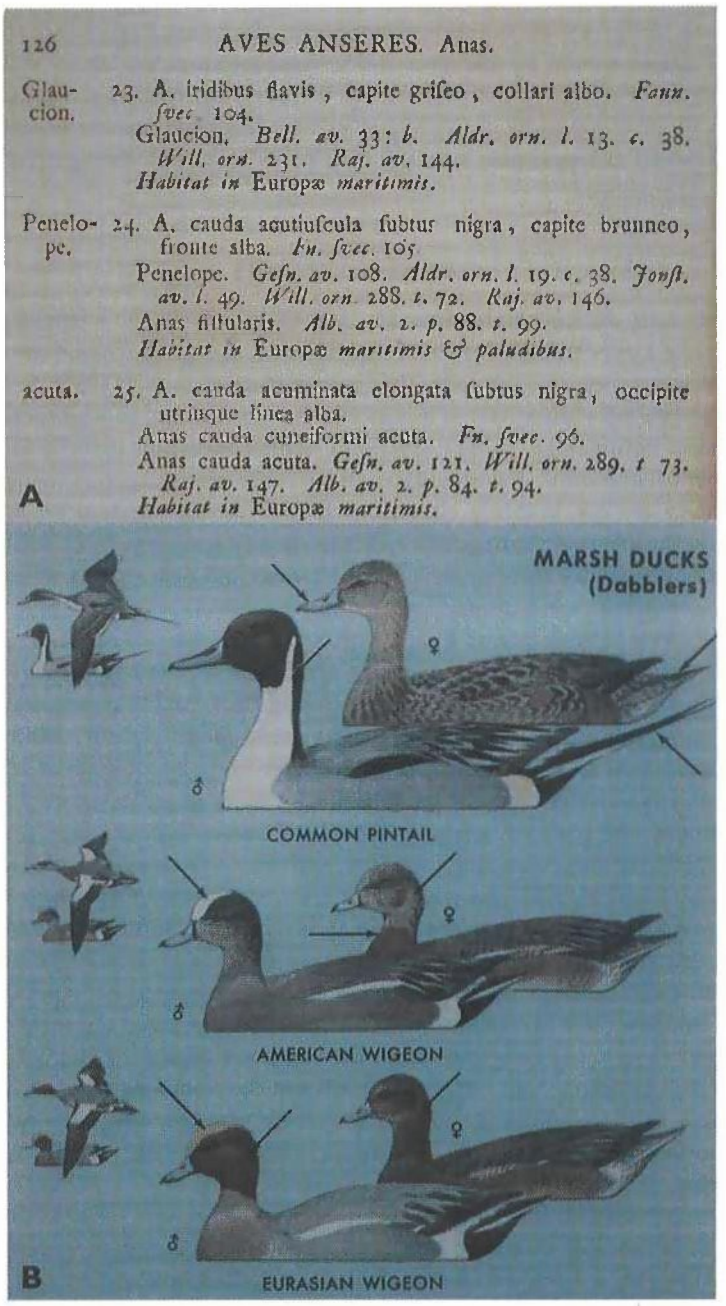


FIGURE 13-2 Examples of identification aids. A. *Systema Naturae* (1758); B. Peterson's Field Guide to the Birds (1980).

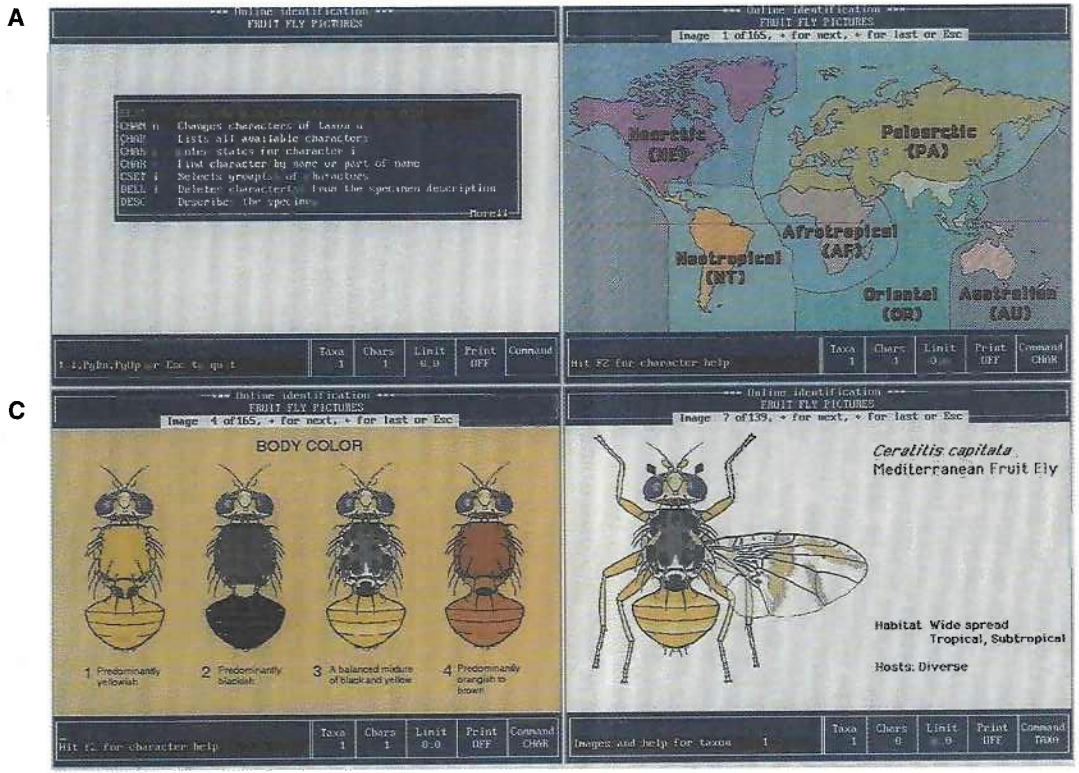


FIGURE 13-3 Computer screens from the Fruit Fly Expert System. A. Menu for commands; B. Character image for biogeographic regions; C. Character image for body color; D. Taxon image for Mediterranean fruit fly, *Ceratitis capitata*.

eliminated immediately by restricting the data set according to geographic location (Figure 13-3B) or other biological data. Any character can be chosen in any order, or the computer can list the best characters based on their ability to separate the remaining taxa. Characters are illustrated and multiple states are allowed (Figure 13-3C). This speeds the identification process in two ways, by enabling direct comparison of images with the specimen and by reducing the total number of decisions that must be made, because more than the traditional two alternatives can be evaluated efficiently at one time. Characters and computer commands are explained in help files that can be accessed at any time. Computer commands are either selected from menus or entered directly (Figure 13-3A). How closely a specimen must match characters (the error limit) can be set, so more matches must be made before a taxon is rejected. Errors, once detected, can be corrected easily without stepping through all the characters again. The identification can be verified easily. The computer can generate a complete description and image of any taxon (Figure 13-3D), list only the differences between the specimen and another taxon or between any two taxa, or generate a list of all the diagnostic characters for a particular taxon. The Expert System is not a panacea: unusual specimens, those outside the domain of the data set or with distorted features, still will have to be sent to systematists. Systematists are still needed, but by relieving them of most routine identifications, they can be more productive, and users will get their identifications faster.

CONCLUSION

The Systematic Entomology Laboratory has been and is committed to the characterization of biodiversity to help society develop an understanding of biodiversity and the ability to use it wisely. We are building classifications, a set of keys to enable us to better communicate about biodiversity. We also are committed to developing tools, such as expert systems and biosystematic information databases, to allow users to obtain these keys (names) and to know what are the best keys (valid names).

REFERENCES

Bisby, F. 1994. Global master species databases and biodiversity. Biol. Int. 29:33-38.

Hawksworth, D. L. 1994. Developing the bionomenclatural base crucial to biodiversity programmes. Biol. Int. 29:24-32.

International Botanical Congress. 1994. International Code of Botanical Nomenclature. Regnum Vegetabile 131. Koeltz Scientific Books, Königstein, Germany. 328 pp.

International Commission on Zoological Nomenclature. 1985. International Code of Zoological Nomenclature, third ed. International Trust for Zoological Nomenclature, London. 338 pp.

Linnaeus, C. 1758. Systema Naturae, tenth ed. L. Salvii, Stockholm. 824 pp.

Miller, S. E. 1992. Specimen databases and the lack of standard nomenclature: A proposal for North American insects. Insect Coll. News 7:7-8.

Pankhurst, R. J. 1991. Practical Taxonomic Computing. Cambridge University Press, Cambridge, England. 202 pp.

Peterson, R. T. 1980. A Field Guide to the Birds. A Completely New Guide to All the Birds of Eastern and Central North America. Houghton Mifflin, Boston. 384 pp.

Ride, W. D. L. 1991. Justice for the living. A review of bacteriological and zoological initiatives in nomenclature. Pp. 105-122 in D. L. Hawksworth, ed., Improving the Stability of Names: Needs and Options. Regnum Vegetabile 123. Koeltz Scientific Books, Königstein, Germany. 358 pp.

Sneath, P. H. A. 1986. Nomenclature of Bacteria. Pp. 36-48 in W. D. L. Ride and T. Younès, eds., Biological Nomenclature Today. IRL Press, Eynsham, England. 70 pp.

Sneath, P. H. A., ed. 1992. International Code of Nomenclature for Bacteria. 1990 Revision. American Society for Microbiology, Washington, D.C. 232 pp.

Stearn, W. T. 1957. An introduction to the *Species Plantarum* and cognate botanical works of Carl Linnaeus. In C. Linnaeus, Species Plantarum (a facsimile of the first edition of 1753). The Ray Society, London. 176 pp.

Thompson, F. C., A. L. Norrbom, L. E. Carroll and I. M. White. 1993. The fruit fly biosystematic information database. Pp. 3-7 in M. Aluja and P. Liedo, eds., Fruit Flies: Biology and Management. Springer-Verlag, N.Y.

World Conservation Monitoring Centre. 1992. Global Biodiversity. Status of the Earth's Living Resources. Chapman and Hall, London. 585 pp.