The Fruit Fly Biosystematic Information Data Base

F.C. Thompson, A.L. Norrbom, L.E. Carroll, and I.M. White

Introduction

Biosystematic information is all data that may be useful to man about organisms, such as what is it, what is it called, what does it look like, where does it occur, what does it do, when does it do it, and what does all this mean to me (= economic importance). Biosystematic information is organized by names, arranged in a hierarchical classification based on shared (synapomorphic) similarities. Hence, biosystematic information can be obtained with a name, a species, or names of associated objects. Names are obtained by identifications, and identifications are made by matching attributes of unknown with known organisms. While that is all a logical sequence, what is reality? In reality, most users obtain biosystematic information merely by asking systematists for it!

The Systematic Entomology Laboratory has been America's primary source of Biosystematic Information about insects and mites for more than a century. The information provided has been a critical component to our agricultural success. Today increased concern for preservation of biotic diversity and environmental quality along with the traditional needs of agriculture has increased the demand for Biosystematic Information, while fiscal problems have seriously eroded the resources available to meet the demand. One merely needs to note that over the last few years the Systematic Entomology Laboratory has lost 4 scientific positions and 7 technical support positions, while the number of identification requests has been stable. These statistics clearly document the problem. As additional base funding is unlikely given the current deficit, the Systematic Entomology Laboratory must explore new technologies for solutions to this crisis. The situation is similar at all other major centers for systematics (Australia (CSIRO), United Kingdom (BMNH), Canada (CNC), etc.).

Both raw data and users' inquiries are funnelled to the systematists, who answer the inquiries, provide identifications, and compile and synthesize

From *Fruit Flies: Biology and Management*, edited by M. Aluja and P. Liedo, Springer-Verlag New York, Inc., 1993.

biosystematic data in technical publications. Thus, every inquiry goes to the systematists and the answers must return from them. Thus, the systematists are the classic "bottle-neck" in the flow of biosystematic information.

Given the flow of data, the obvious target to concentrate on is the systematist. Relieve the demands on the systematists and increase their productivity and the flow of biosystematic information will increase. Demand can be reduced if the users' needs can be answered directly. Thus, providing more direct access to Biosystematic Information in a format that users can understand and work with is one critical task. Productivity can be increased by reducing redundant data handling and by making data sharing more efficient. We can't waste valuable manpower, so literally every keystroke must be preserved and shared so together the diminished few can do what once many did and now every one wants! A Biosystematic Information system is therefore proposed as a major contributor to resolving the crisis in Biosystematics.

A Biosystematic Information system could do much to resolve the Biological Diversity crisis. The simple integration of expert systems, relational databases, and image processing provides tools to make the systematics more efficient, store and integrate the knowledge of the systematists (experts), and provide that knowledge to users as needed. Systematists will be more efficient, more productive, and have more freedom to pursue critical research. Users will get more immediate access to more information and will have more independence. Biosystematic Information systems (Figure 1), will divert the demand for information from the systematists and increase the rate of flow of information through the system. Additionally, automation will eliminate redundant data handling, maximize data sharing, increase the rate of data processing and reduce the cost and bulk of data storage.

The core of a Biosystematic Information system is a relational database. Because Biosystematic Information is already organized hierarchically, a simple relational database model can be generated for it (Figure 2). The data needed for a Biosystematic Information system would be assimilated by systematists as they worked, and when a sufficient amount was accumulated, that data would be automatically formatted and distributed to the users to be used on their microcomputers. CD-ROM (Compact Disk-Read Only Memory) would provide the capacity to store immense amounts of data. On CD-ROM, regular text based data could be stored with images and sound-recordings. A single 5 1/4 inch CD-ROM disk could contain more than 15,000 pictures of fruit flies and some 150,000 pages of information about them. For users, an expert system interface would be developed so that the technical data would be presented in a "user friendly" manner.

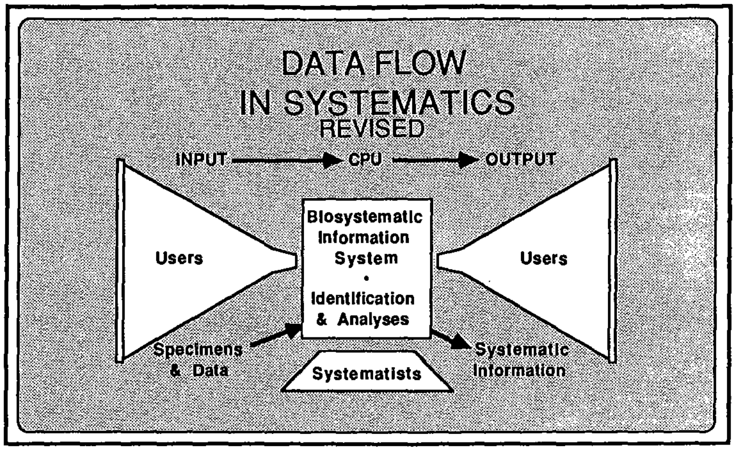

Figure 1. Data flow in systematics.

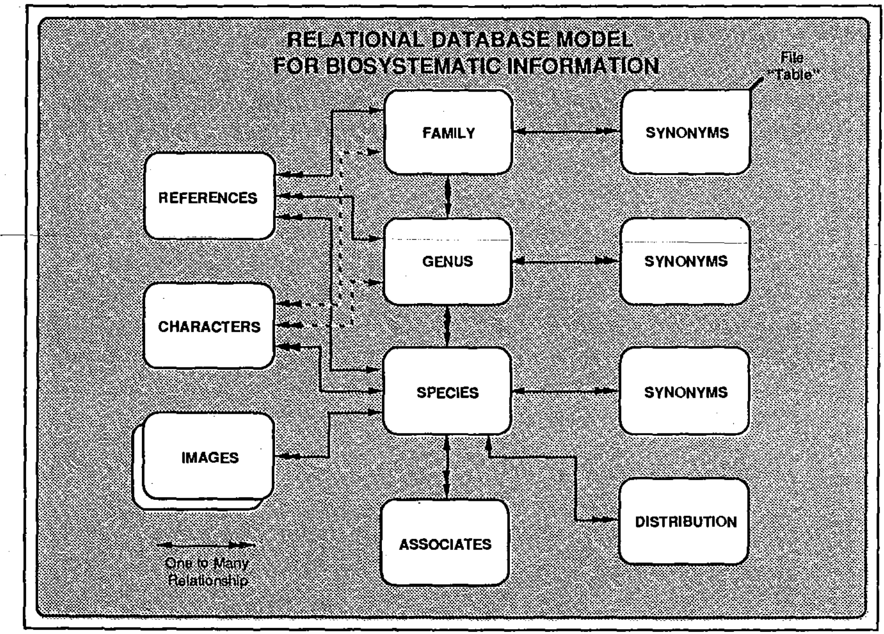

Figure 2. Relational Database model for Biosystematic Information.

Goals of the Fruit Fly Project

The Fruit Fly Biosystematic Information Databases is a prototype of this system that is being developed at the Systematic Entomology Laboratory. Continually updated in a relational database are basic nomenclatural, host, parasite, and distributional information for the species of Tephritidae of the world.

A user-friendly Expert System, ONLIN6, is being developed in collaboration with Dr. Richard Pankhurst. This will provide users, such as APHIS-PPQ identifiers, with an illustrated interactive key which will produce the specimen identification necessary to access related information in the Database.

The Database will be published in hard copy in the traditional catalog format, but also the Database and Expert System will continue to be updated and will be made available in CD-ROM, which users can use on their own computers.

Current Status

Data entry for the Fruit Fly Biosystematic Information Database is approximately one-third complete. The nomenclatural portion of the Database is nearly completed and is in the proofreading stage. It currently contains 794 generic names (600 valid) and 4951 specific names (4090 valid). We are now working on the type data, including type depositories. Completion of the host section is anticipated in 1991.

Our working Expert System prototype currently contains morphological data in standard DELTA (DEscriptive Language for TAxonomy) format for more than 130 characters for 14 species representing the major subfamilies and tribes of Tephritidae, as well as several of the most important pest species. Integration of additional data for 54 species of the major pest group, Dacinae, is anticipated in 1991, and our eventual goals is approximately 200 species. Images for illustrations are being prepared, and the software is undergoing further development. Host and distribution data will also be incorporated in the same format as morphological data.

How the Expert System Works

Traditional taxonomic keys function like primitive expert systems, but with an incomplete data matrix. The earliest keys were just text, with relatively few characters. Keys have been improved over the years by adding more characters, as well as illustrations. But the main problem with keys is inflexibility. There is a set pathway (often a long one) through the key to each species; a single mistake may lead to an erroneous identification; a single missing character may leave the user at a dead end. Verification of the identification requires reading complete descriptions to find all the characters to check.

An expert system is much more flexible. We can eliminate many taxa immediately by restricting the data set according to geographic location or host data. We can choose any character of any sex or stage in the life cycle, in any order which seems best to us. We can ask the computer to select the best characters for us, based on their ability to separate the remaining taxa under consideration.

Characters are accompanied by illustrations, and multiple state are allowed. This speeds up the identification process in two ways: by enabling direct comparison of images with the specimen (rather than reading text), and by reducing the total number of decisions which must be made, because more than the traditional two possibilities can be efficiently evaluated at one time.

Characters are also accompanied by help files which can be accessed at any time. Even so, we admit the possibility of error (e.g., a poor or aberrant specimen) and can ask the computer to tolerate an error or two before rejecting a taxon. Errors, once detected, can be corrected easily, without stepping through all characters again.

The verification process is also much easier. Although we can obtain a complete description, just as in traditional taxonomic references, we can also ask the computer to list only the differences between our specimen and another taxon, or between any two taxa. Or we can ask for a list of all the diagnostic characters for a particular taxon. With the identification process complete, we can now query the Database for complete nomenclatural and distributional data, as well as pertinent host and parasite associations.

We realize that the Biosystematic Information Database will not be a panacea. Unusual specimens, those outside the domain of the Expert System or with distorted features will still have to be sent to the systematist. We still need more taxonomists, and identifiers will need some training in the use of the system. But by relieving taxonomists of most routine identifications, they will be more productive in research and in providing additional biosystematic information.