## NEED FOR THE BioSystematic Database of World Diptera

The BioSystematic Database of World Diptera is a resource to ensure knowledge about flies (Diptera) is easily accessible to all as the BDWD provides a framework to index and integrate current and future information.

Since the establishment of the Convention on Biological Diversity (1994 [http://www.biodiv.org/]), nations and people everywhere are aware of the importance of the life around us. Much has been written in both the scientific and popular media. The world biota is disappearing at an alarming rate never before equaled in the history of this universe (Wilson 1992). That is not disputed. What we do not really know is what is being lost as our knowledge of most organisms is abysmal. Lists of birds and mammals are available, but for the little creatures, such as insects, no comprehensive inventories exist. Hence, the Convention and many others have declared that an inventory of what is known is of critical importance and immediate priority. New organizations, such as the Global Taxonomic Initiative of the CBD (GTI 2000), Species2000 [http://www.sp2000.org/], Integrated Taxonomic Information System (ITIS [http://www.itis.usda.gov/]), Global Biodiveristy Information Facility (GBIF [http://www.gbif.org/]), have arose to address the problem, but all are dependent on scientists to find the resources and time to do the work. The BDWD exists to fill the need for flies, a group that includes more than 10% of the known (described) biodiversity of the world. The basic system is built, specialists are ready to fill it with information, so all can access our knowledge about flies. With additional funding the BDWD can be quickly completed, delivering all the names for some 150,000 fly species, more than 10% of the known biodiversity of the World. No other large block of names is yet ready to be delivered.

Flies, midges, gnats, bots and other two-winged insects represent a major clade of life (order Diptera). They are found in almost all habitats, although rare in marine ones and absent from high polar ones. While some dipterans are as beautiful as birds and have attracted the attention of amateurs, many others are of more critical importance to man as vectors of diseases, pests of food and fiber, model systems for research (*Drosophila*), precise bio-indicators for conservation and environmental quality assessment (Chironomidae), pollinators of flowers and biological control agents against weeds and pests. These interests have led specialists on flies to cooperate to produce regional catalogs (Stone el alia 1965, Papavero 1966- Crosskey1980, Delfinado & Hardy 1973 -79 ; Evenhuis 1990, Soos & Papp 1994-), specialized catalogs (Evenhuis 1994 (fossils), Sabrosky (1999, family-group names), and now manuals (McAlpine 1981-9, Papp & Darvas 1998-2000). Unfortunately, these catalogs are out of date and not easily accessible. The BDWD provides a means to revise and up-date these regional catalogs, as well as access the information through a single comprehensive index. For the general user the BDWD will allow access to information by various species attributes such as distributions (alien, invasive, native), economic importance (pest, beneficial, endangered), associates (hosts) and habitat (aquatic, marine, terrestrial). Lastly, the BDWD will serve as an internet portal, providing links to other informative pages elsewhere on the World-Wide-Web.

The BDWD essentially consists of two components, a nomenclator and a species database. Also a number of tools are included as part of the system to aid specialists in building and

maintaining the BDWD. The nomenclator allows users to find the single correct name for each dipteran as well as critical nomenclatural and taxonomic information about all names. To access information through the nomenclator, one merely enters a name or even just part of a name. Searches can also be made by the author or authors of names, the year in which they were established, etc. The nomenclator returns a list of all names which match the criteria supplied or a statement that the name is not recognized. Various fuzzy logic alternatives will be explored and incorporated to help the user identify spelling errors. One then can select individual names listed to view the full details available on each name. These details include the full name, the correct name, the type and source of name, and quality assessment of and authority for the information. If the name is not found by the system, the user will be requested to send the name and the context in which it was found to the BDWD project. This user feedback will allow the project to identify errors. The species database provides the attributes of general interest and the links to other sources of information on the internet. Tools include a directory of specialists working on Diptera, databases of authors, museums and serials. These tools include additional information not found in the nomenclator or database. For example, one can use these tools to determine the best date for the publication of an issue of a serial, dates of birth and death of an author or the location of a particular museum; so these tools are useful to a broader range of users beyond those interested only in flies.

Other nomenclators and species databases are now available online, but none are of the magnitude and comprehensiveness of the BDWD. Nomenclators, such as IPN [http://www.ipni.org/), ZR TRITON (http://www.york.biosis.org/free_resources/ion.html), ITIS [http://www.itis.usda.gov/], provide only minimal information and never data on types of the names. The few species databases currently online, none of which cover flies, are fragmented and limited, requiring a prior knowledge of the classification to use, and not allowing retrieval by species attributes nor obsolete names (for example, Orthoptera Online Species file [http://viceroy.eeb.uconn.edu/odb/pages/main.html]).

The taxonomic breath of the BDWD is all taxa (members) of the order (clade) Diptera, recent and ancient. The nomenclator will include not only all scientific names, but misspelling and significant misidentification. The geographic scale is global; time scale is all names since 1758, the starting date of Zoological Nomenclature. The BDWD will be the first comprehensive coverage of flies since the *Systema Antliatorum* (Fabricius 1805).

The BDWD is more than an idea. The project, databases and WWW interface have been created. Data capture is nearly 2/3 complete, with approximately 160,000 names already entered, including all family group (4,324) and genus group (21,084) names. Almost all names for the Nearctic, Neotropical and Afrotropical biotic regions have been entered,and all the names for 106 families have been completed. Altogether, some 130,000 species-group names have been captured. We estimate some 40,000 remain to be captured. So far, the names of the true fruit flies (Tephritidae) and solider flies (Stratiomyidae) have been peer-reviewed and certified as authoritative. A number of specialists have name data in various electronic formats, from word-processing files to databases, and are willing to contribute these data to the project. The current status of the BDWD can be easily checked by visiting our WWW Diptera site (www.diptera.org) and browsing the various pages under Names.

## MANAGEMENT PLAN

### Strategy

Much data about scientific names exists. Beyond the name is its history (who created it and who subsequently used it), typification (what is its type? who designated the type? where is the type?), its placement within a classification, and its underlying concept now and in the past (what characters circumscribe the concept and what are their histories). Magnitude is the challenge for entomology: There are probably 1.5 million different kinds of flies; can entomologists know them all? The key to biosystematic knowledge is names: All information is indexed by scientific names. Thus, the first step to knowing biodiversity is to have a stable and comprehensive naming system. To develop this system, we need data about names. To know flies, we need information about their names and about the flies themselves. How does one start building such a comprehensive information system? Given limited time, there must be a trade-off between depth of information gathered and breadth of taxonomic scope.

Within the next few years, we could capture a lot of data about a few flies or the minimal critical amount about of all of them. Much of the data about names (typification, circumscription, detailed taxonomic placement, history) is useful only to the systematist, who must make decisions about names and their underlying concepts. Users of names depend on systematists to know these data and to make correct decisions based on them. Users want only to know what the correct (valid) name is and the basic classification of the name (Class, order, family & genus). So, we will concentrate on providing the basic data that users want first. This information will be provided by the Nomenclator component. After that we will concentrate on the species database and portal. To ensure that our concept of the species database and portal are useful, we will give high priority to those thousand or so species of major importance to man. For the vast majority of described flies, the only information available is what it looks like and when and where a few specimens were collected. Fortunately, that species information is included in both the nomenclator and the species database.

Our basic strategy is to capture the critical data from the best available sources (see below). Once this is done, the working records are available to the community as a whole. However, these working records will then be distributed to key specialists to be checked against original sources and to have the taxonomy brought up-to-date. When the specialists have revised these working records, their efforts will be peer-reviewed. After the records have been reviewed, they will be certified as authoritative (see BDWD Quality Assurance statement [http://www.sel.barc.usda.gov/diptera/names/bdwdqa.htm]).

### Data Model and Dictionary

A definitive data model for all nomenclatural and systematic information has yet to be developed and adopted. However, all data models published (ASC 1992, IOPI 1993, ECN 1990, CSIRO 1993) as well as ones being developed (e.g., Beach and Blum & Humphries) have a minimal set of common data elements. We capture data for the critical subset of these common data elements. Our data model and dictionary, hence, are compatible with all models we know and our data can be imported easily into any database system based on any community accepted model. Given the trade-off of depth and breath (see above) and the lack of a detailed and comprehensive data model, the logical option for the present is **to capture and validate the basic data for minimal critical common elements**. This is our strategy;

the details are presented below. All community data standards will be followed. First and foremost, the *International Code of Zoological Nomenclature* will used. Co-editor (Thompson) was a member of the Commission and its Editorial Committee. BIOSIS (*Zoological Record*) standards for the treatment of author names and other data not covered by the Code will be followed.

The BDWD consists of seven main tables: 1) species-group names, 2) genus-group names, 3) family-group names, 4) references, 5) species, 6) distribution, and 7) associates. The six tables can be used separately or in various relational models depending on the problem or query being asked. The nomenclator, for example, is a special database built from the three names tables and optimized for fast WWW interactions. The relationship among these main tables and the WWW user-interface are diagramed (...). Details about the fields and the standards for their contents are available at the WWW Diptera site in a downloadable Adobe pdf file [http://www.sel.barc.usda.gov/diptera/names/bdwdstan.pdf]. The basic contents are as follows:

The names tables (family-group, genus-group and species-group) contain all the nomenclatorial details (name, author, year, page, type and associated data, biotic region, status, valid name, valid name author, etc.) necessary to solve any name problem broken down into separate fields. These tables are linked to each other in a hierarchy. Each record is also linked to the reference table which contains details about the published works that contain the details. Other support tables are museums table containing the full names and addresses for type-depositories and the status table containing the full explanation of the status codes. The nomenclatorial data is broken into three tables because each category of names has slightly different data components for types, valid names, etc.

The reference table contains all essential information about published works. The current structure is rather simple, and will be normalized over the course of the project. Currently the fields are author, year and title. Both the author and title field will be parsed to their critical elements and the author will be linked to a worker table.

A species table will be built to contain more attributes about species, such as their economic importance (pest, biological control agent, pollinator, indicator, etc.). This table will be linked to the distribution and associates tables and provide a URL link to species pages elsewhere on the world-wide web. Future enhancements will be to normalize the URL link and add other more traditionally published citations into a separate child table (citations).

The distribution table will include data about locations where the species is known to occur (biotic region, country, major component of large countries (such as states for the USA), status within location (native, endemic, alien and invasive). This will be a joint table between the species table and a table of standardized political and biogeographic units. Besides the two foreign keys, information on the status within those units (endemic, native, alien/introduced, invasive) and appropriate comment will be included.

The associates table will include data about other species associated with the named one, such as plant or animal hosts, flowers pollinated, parasites, predators, etc. This will also be a joint table between the species table and a table of standardized biological roles (predator, prey, host, et cetera). The record will also have fields to further qualify and document the role. These data can

be extensive for some species; hence, the extent to which the associates table will be populated will be determined by the contributing specialists.

A number of tools have also been developed to aid workers in bulding the BDWD. The most important is the directory of dipterists (now online), which allows users to know who is doing what on flies. Other tools include databases of serials and their publication dates, workers (past and present), type-designations, historical place names, museums, et cetera. These tools ensure that consistence in the contents of various database, but also provide useful information directly. For example, the museums database provides standard acronyms used for the type-depositories in the species-name database, also providing names, addresses, contact people, et cetera, for these museums.

FileMakerPro was selected as the platform to be used as it is inexpensive, runs interchangeable on Windows and MacIntosh computers, and includes a WWW interface. Previously some of the data of the BDWD were on a Wang VS minicomputer system. We were successful in migrating those data. As the data are maintained in a relational structure, we are confident that any future migrations to other platforms will also be easily made.

MANAGEMENT PLAN

The organization of the project is three tiered. The core are the co-editors, Neal Evenhuis and F. Christian Thompson (Bishop Museum & Systematic Entomology Laboratory. They make decisions on database structure, set data and editorial protocols, and coordinate with outside contributors to the project. Together they are responsible for the construction of the databases, their completion and integrity. Their institutions have assumed responsibility for the maintenance of the BDWD, updating and dissemination of the completed product.

The second group is an Editorial Committee. The members are selected to represent diverse interests by scientific study, biological group, and geographic location. The current members are Drs. Thomas Pape (Swedish Museum of Natural History, Stockholm), Adrian Pont (University Museum, Oxford), and Hiroshi Shima (Kyushu University, Fukuoka). They are available to advise the steering committee. The editorial group is now small, but as more products, such as the species database, comes online, the group will be expanded to more fully represent of all users.

The third level of the organization consists of a network of collaborators. The list continues to grow as the project becomes better established. Responsibilities of collaborators vary. Some will provide parts of the database relevant to groups they work on. Others will coordinate other collaborators for major groups as well as provide data. Lastly some will act as peers to review the contribution of others. However, the key task of all specialists will be to ensure that the nomenclatorial and taxonomic information conforms to the highest community standards. More information is available on the BDWD team (see our people page [http://www.sel.barc.usda.gov/diptera/names/bdwdpeop.htm]).

The management plan incorporates 5 levels of activity. Advice will be sought from the Editorial

Committee and scientific community through the International Congresses of Dipterology. For example, this project description and the comments it generates will be shared with our advisors. Advice will be considered by the editors and incorporated into the project. The databases, standards and protocols established by the editors will be implemented as they will oversee the work of catalogers, handle data flow to and from collaborators, and insure the integrity of the information. Data will be searched for and/or entered by the catalogers, who will make modifications as directed by the editors or in accordance to standards and protocols. Data entry has been and will continued to be done by USDA personnel. The master copies of all of databases will be maintained in Washington. Periodically versions will be generated and posted on the World-Wide-Web (target quarterly updates) and archive copies of all databases on CD-ROM will be disseminated to all key workers. Annually the databases will also be disseminated in the *Diptera Data Dissemination Disk* (A CD-ROM journal). Names and associated data for family or smaller units will be sent to systematists, who will revise the information. The revised information will then be peer-reviewed and certified.

Tasks

The principal tasks to be completed are five: 1) Data entry of names not yet captured; 2) Building a bibliography and linking those references to name and species records; 3) Basic editing, review and certification of names already in the BDWD; 4) Enhancing the BDWD to include a species database and portal, better online tools and editing features as well as a streamline conversion to traditional printed format; and finally, 5) Maintenance of the BDWD through the addition of new names and references as well as the periodic re-certification of the taxonomic status of all names.

The first task is likely to be completed within the next two years at current funding levels. Building a basic bibliography is likely to be complete over the next couple of years as well the task of linking the records together. A quality review of the bibliography with the addition of critical data elements not usually found in traditional printed bibliographies, such as precise publication dates, will not be done as quickly. This and the task of review and certification are strongly dependent on abilities, availability, support available for, et cetera, of specialists. Specialists represent the critical bottleneck to the completion of the BDWD. With few and fewer active specialists, and the remaining specialists with many conflicting priorities and little or no technical support, the key challenge for the BDWD is incentives to specialists so the BDWD is their highest priority.

Enhancing the capabilities and features of the BDWD is critical. A completed nomenclator will be a useful tool for specialists who understand systematics and nomenclature, but other users will find it of limited use. So, the challenge is to build a species database and portal for general users (non-specialists) to use. These users will want to retrieve information by attributes of species, such as where they occur and what they do. For example, all the pests of corn that occur in Chile is a typical query that a user may make of the species database. The design and testing of the species database and portal will take a year or two and depends on the infusion of new funding. Online tools need to be developed to assist specialists and editors, to eliminate redundant work, and to ensure consistent, uniform values in many data fields. For example, in handling type-locality data, specialists will find the same locality under different names and sometimes in

different countries. To address this problem a working database of historical localities will be develop to ensure only one name will be used for each location. Another tool will be a database of the precise publication dates (to the exact day if possible) for various journals. As these tools are necessary for our specialists and may be useful for specialists working on other organisms, priority will be given to building them and placing them on WWW. Currently the BDWD is being built by specialists working independently either by using word-processors or their own databases, with their results being manually merged back into a single master database. For security and for easier mass editing operations this is an effective procedure, but for the long-term an online editing interface to the BDWD must be buily so specialists can revise their records directly. With new funding this interface will be a high priority.

Once properly built, the long-term maintenance of the BDWD will be the only remaining task. Given the present rate at which new taxa are described and old ones revised, maintenance will not be a large task. Only a few thousand taxonomic actions per year are now being listed by the Zoological Record for Diptera; these changes could be sent directly and in digital format to the BDWD for a fee. The changes would then be available for specialist review and certification in the online version of BDWD, and given availability of the online editing interface the specialist would do this directly. Hence, the low level of maintenance required by the BDWD in its future format could be easily assumed by the Systematic Entomology Laboratory. Obviously major conversion required by the inevitable evolution of computer hardware and software will require additional support.

Specific Tasks for the Collaborators

The principal tasks required of collaborators are as follows. 1) Provide a higher classification for their family-level taxon. That is, a list showing the placement of each genus within their appropriate taxa, such as tribe and subfamily. 2) Verify genus-group name data, reviewing: a) distribution of valid genera (to Biotic Regions) and b) current taxonomic status. 3) Revision of species and species-group names, providing: a) distribution data for species; b) other species attributes as known; c) taxonomic status of species-group names; d) addition of obsolete combinations as known; e) source of synonymy of invalid names as known; f) type information (status of type, kind of type, location of type, and additional type information as needed (source of lectotype designations, for example)). 4) Preparation of introduction to their family-level taxon.

Data Sources and Previous Work

Catalogs of flies exist for various biotic regions (Nearctic (Stone, et alia, 1965), Neotropical (Papaver 1966-), Afrotropical (Crosskey 1980), Oriental (Hardy & Delfinado 1973 -79), Australian/Oceania (Evenhuis 1990), as well as fossil flies (Evenhuis 1994). Beyond the regional Diptera catalogs, various specialists have published family-level treatments and Sabrosky (1999) has treated all family-group names. Other comprehensive sources of zoological names are Neave (1939- , genus-group), Sherborn (1902-32, species-group until 1850), and the Zoological Record (1864- ) (details about these sources can be found in the sources document [http://www.sel.barc.usda.gov/diptera/names/bdwdsour.pdf]). Clearly the task of assembling a

database of World Diptera is feasible. The differences between this project and prior work are magnitude and inclusiveness.

The knowledge needed to build the catalog is today scattered across numerous published works, most of which are outdated, and in personal research files. Some of the data are summarized in catalogs, a few of which are based on computerized databases. More or less complete taxonomic catalogs exist in some form for most taxa . Many of these are even in some kind of electronic storage. Many of them, however, are not generally available and all are not quite complete, but these problems are solved with this project. Only specialists know where to find **all** of the names and know how to treat them, so that they conform to the rules of zoological nomenclature. This project seeks funding to work with the specialists, to consolidate their data, and to disseminate the resultant information to users.

The status of our knowledge of the world flies was recently summarized (Thompson 2000) and is documented on the status summary page [http://www.sel.barc.usda.gov/diptera/names/bdwdss.htm] and statistics table [http://www.sel.barc.usda.gov/diptera/names/bdwdstat.htm].

Sponsors and Current Support

The BDWD is sponsored by the USDA Systematic Entomology Laboratory and the Bishop Museum. The project has been endorsed by the International Congresses of the Dipterology, a scientific member of the International Union of Biological Sciences and participates with the Species2000 program and the Integrated Taxonomic Information System. Both these programs already contain a significant number of data records which originated with the BDWD (about 17% names in ITIS are from the BDWD and about 90% names in Species2000 are from ITIS (or about 15% from BDWD via ITIS).

The Systematic Entomology Laboratory has supported the BDWD by providing 10% of one research scientist (the Co-PI) and 25% of one data entry clerk as well as the necessary computer support to maintain the Diptera WWW site. Part of this site was built with support from NSF PEET grant to Wayne Mathis (DEB #95-21773). In 1999 the Schlinger Foundation provided a small grant to build the WWW interface to the BDWD and to support a post-doctoral fellowship to work up the Lauxanioidea families. At the current rate of support from USDA alone, the BDWD will be completed some time around 2010.

The BDWD is not a USDA project nor even a Bishop Museum one. The BDWD is a community-based project under the leadership of these institutions. Many dipterists world-wide are committed to the BDWD project;, support is necessary to coordinate, assimilate and disseminate the work of the Diptera community. With its mandate to do fundamental research on organisms affecting US Agriculture, USDA has contributed significant support to the BDWD and will continue to do so. Unfortunately with less than 10 % of flies being of significant agricultural importance, USDA can not justify supporting all the BDWD needs. Thus we are seeking additional funding for the accelerated completion of the BDWD. So, the legitimate questions are: 1) is the BDWD a useful product for Society and Science as infrastructure and as a model for

other megadiverse taxa; and 2) given what has been achieved, are the current workers and plans sufficient to achieve the desired goals. We believe the answers are yes, and, therefore, other funding sources (such as NSF) should contribute a share to see an accelerated completion of the BDWD.

Electronic Products

The electronic products are the nomenclator and species database available on the WWW (Internet) and the source databases available in the digital journal, the *Diptera Data Dissemination Disk* (a CD-ROM scientific serial published by the North American Dipterists' Society), as well as various tools. The nomenclator and species database are specialized databases optimized for the WWW (Internet) and are derived from a number of source databases. These products are available today, although enhancements and more complete data coverage are planned for the future. All technical information (data model and dictionary, standards, protocols, etc.) are also available here at our WWW site (www.diptera.org). The Systematic Entomology Laboratory is committed to maintain the BDWD as long as it remains a useful product to the scientific community. The information is annually archived by its publication in the *Diptera Data Dissemination Disk.*

Copyrights

The BDWD operates under US Law, especially the fair-use provisions of the Copyright Law. As the major supporter of the project is an US Government agency, the BDWD is without copyright. The data and information remains freely available as noted above. Collaborators will be asked to transfer limited rights to the BDWD for the free scientific and non-commercial use of their contributions.

Value-Added

We have and will continue to train students, who have and will contribute names and species information to the BDWD. FCT has served as advisor to two PEET grants, where the nomenclature and species data has been generated for incorporated into the BDWD.

Scientific Context

To conserve biodiversity, society must not only know its magnitude but also the particularities of each component of that biodiversity. Time and again scientists have demonstrated that great benefit can be derived from better knowledge of biodiversity. However, the key to sustainable use of biodiversity comes from being able to associate prior knowledge with current information, to communicate precisely about the potentialities of discoveries, and to define the scope of related sources of similar or greater benefits. All these attributes result from properly assigning a name to an organism and, thereby, placing the organism within a phylogenetic (predictive) classification (Janzen 1991, Thompson 1996).

The BDWD will provide a resource for these essential and initial steps. While assignment of an

appropriate name is essentially an identification, which is not directly addressed by the BDWD, the BDWD can provide useful information to that process. And when the identification is made, the BDWD is an essential resource to ensure that proper name is attained. That is, although an identification tool (such as a taxonomic key) may be useful, its nomenclature may be obsolete. The BDWD allows users to find the current valid names from an obsolete one. Also, the BDWD with its species database may also be useful when appropriate identification aids are not available, but some species attributes are known. Such questions may appear as an unknown leaf-mining pest of soybean being discovered in the USA. The BDWD with its ability to support queries by species attributes, such as host associations, would allow for a quick and easily restriction of the possibilities to a few or one fly species.

The availability of a complete BDWD may provide scientists with the data to answer questions about scientific productivity, estimated magnitude of unknown biodiversity, changes in diversity over time, etc. Everyone estimates that there are many more insects than presently described. Those estimates then beg the question of the feasibility and cost of describing all that biodiversity. To estimate the future, the trends of the past are critical. The BDWD data would allow the determination of the average (as well as high and low extremes) of productivity of systematists. The BDWD would provide data for the estimation of the total size of the Diptera clade from species description accumulation curves or other measures of species richness (CHAO2, etc.). Knowing about what fossil and recent species have been described can allow estimation of fauna turn-over, extinction, etc.

**In summary**, a complete BDWD will be a useful resource for vast array of other scientists, regulators, students, and the general public.

Literature cited

Most of the references cited as sources of Diptera names and names generally will be found on our sources page [http://www.sel.barc.usda.gov/diptera/names/bdwdsour.pdf]. Only the general printed references are as follows:

Fabricius, J. C. 1805. Systema antliatorum secundum ordines, genera, species. C. Reichard, Brunsvigae [=Brunswick]

Janzen, D. H. 1991. How to save tropical biodiversity. American Entomologist 37: 159-171.

Thompson, F. C. 1996. Names: The keys to Biodiversity. Pp. 199-211. In Reaka-Kudla, M. L., Wilson, D. E. & Wilson, E. O. (eds.), Biodiversity II.. J. Henry Press, Washington

Wilson, E. O. 1992. The diversity of life. 424 pp. Harvard University Press, Cambridge.