Automatic Data Processing for Systematic Entomology

Biosystematic information is critical for today's world. Every major concern, such as global warming, food supply, environmental quality, etc., has a biological component that is dependent in part on biosystematic information. What is biosystematic information? 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 an hierarchical classification based on shared (synapomorphic) similiarities. Hence, biosystematic information can be obtained with a name. Names are obtained by identification of specimens, and identifications are made by matching attributes of unknown with known organisms. While everyone makes some identifications, for diverse and little known organisms, such as insects, identifications are made by systematists. Systematists need the data derived from specimens (and literature) to make the comparsions which lead to identifications, Specimens and their associated literature form collections. So, ultimately the biosystematic information must be derived from systematists and theft collections. And, therefore, the methodology used by systematists to manage theft collections and to produce biosystematic information is critical. Automated Data Processing (ADP) methods hold the promise of greater efficiency, but implementation appears to have caused problems. We, a small working group, met to investigate both the promise and problems, which were summarized by a series of questions. While these questions and our answers follow, our overall conclusion was that the promise was real, but problems were not, being due more to semantics and lack of communication.

Systematic Entomology is at a critical transition. The goals of Systematic Entomology have been the enumeration of arthropod species and illumination of their characters and relationships. Today, the number of arthropod species is estimated in the 30 to 50 million range, with less than 10 percent of them known. This has led some to call for the abandonment of the goal of complete enumeration and the restriction of our work to those groups already well known butterflies and mosquitoes or of critical importance to man's welfare (agricultural pests & beneficials). Others have suggested, instead, that improvements can be made in the way systematists work. Such improvements would increase the rate of progress, making our goals realistic. Automation offers the promise of greater efficiency. For automated data processing technology to be truly useful, data must be shared. Sharing requires that all users understand how data and information are stored. Efficiency increases when common data standards are used, as less effort is required for conversion between different computer environments, less effort is spent on program development and maintenance, training, etc. This report is the first step toward the development and adoption of common ADP standards for Systematic Entomology.

ADP Philosophy, Strategy and goals

What are the goals we seek from ADP for Systematic Entomology? Who are our users (curators? scientists? students? the public?); and what are theft needs? And, therefore, what is our strategy and philosophy?

Our belief is that ADP offers the best promise of aiding systematics in reaching its goals. Our ADP philosophy is to handle data once and when first encountered, to analyse data frequently, and to generate and disseminate information as needed. Our strategy is to encourage all ADP efforts, to work toward common data standards, and to share data and information, Our goals are to increase research productivity, information dissemination, and users' access to and satisfaction with biosystematic information.

> Given the massive data that systematists must handle to generate biosystematic information, the principal goal we seek from ADP is greater efficiency in data processing and sharing. Specimens and their associated data are wanted by systematists for analysis, the resulting information is desired by all. Given the enormous number of arthropod taxa, valuable manpower cannot be wasted. 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!
>
> The basic problem with ADP standards in Systematic Entomology appears to be that of the blind men and the elephant. Various people have use ADP extensively in their work. Each feels that they know precisely what these ADP standards, the "elephant," should be, but each describes the elephant differently. So, the first question is: Is there really one and only one "elephant?" Second, if there is only one "elephant" can all our different views be integrated into a comprehensive description? Third, can each work independently on their part of the "elephant" so that the results can be used by all [that is, is parallel processing desirable?]
>
> A single comprehensive view of the data and information of interest to all is presented and a standard is proposed for the documentation necessary for sharing data and information. While these are preliminary proposals which may need further modifications, we believe their eventual acceptance by systematic entomologists will allow the community to maximize the promise of ADP. As users have different priorities, no one will begin by implementing the full view, and the approaches used to build the complete database will be different. However, acceptance of the comprehensive view and the standards associated with it, should insure that eventually all data and information can be integrated.

A single comprehensive view of the data and information of interest to all is presented and a standard is proposed for the documentation necessary for sharing data and information. Endorsement of this report by the Entomological Collections Network will establish a protocol and begin the acceptance process for ADP standards for Entomological Systematics. The community needs to study this report, providing its comments to the working committee so that a final report can be prepared for adoption by ECN, Systematic Resources Committee of Entomological Society of America and other interested parties. Ultimately, these standards will be used to develop a consensus among biologists as whole.

Building comprehensive systematic databases may start from inventory of collections or the literature, but both approaches are interdependent as one can not be completed without the other. Different funding sources make these different approaches significant. For example, at the National Science Foundation collection-based inventory work is funded by the Biological Research Resources Program, whereas funding for systematic catalogs (literature inventories) is provided by the Systematic Biology Program. Hence, collection-based inventory work is viewed more favorably among its peer than is literature-based inventory work. This is unfortunate as both are fundamental research resources for biologists and should be considered together on their merits for funding.

Inventory goals will vary in respects to classification hierarchy. Minimally inventory data should be accumulated for higher order groups, such as family units. This is critical for the proper management of collections. Maximally users would like inventory information for species units. Literature-based species inventories (catalogs) are necessary for species level inventories of collections as well as being critical resources for other biologists.

> Specimens, which form collections, and their associated data biological, geographic and temporal are the basis from which all biological information is derived. Biological information is disseminated in publications. Modem databases of biosystematic information can be built from the original sources of data specimens in collections or from the sources of information themselves the literature. Unfortunately, some biosystetnatic information is now only preserved in the written word literature because many specimens from which this information was derived were never preserved or have through time become lost. Likewise, no collection is complete, each having only part of the accumulated mass of specimens.
>
> Modem curation practices will generate taxonomic inventories of collections. If collection labels are to be "typed," then those keystrokes should be saved. By typing label data into a computer, the computer can use the same data to generate both labels and inventories. Verification is the next step: Are the label data correct? While the computer can perform some basic data checks, eventually some of the data must be checked against sources which currently are not automated. Names must be checked against authoritative lists catalogs to insure that they are at least spelled correctly. Another check is whether the name is the proper one or an incorrect synonym. If the ancillary name

documentation data are gathered during the verification process, then literature-based inventory can also be built from the curation process (and vice-versa). Ultimately, however, there must be verification of the identification process. The name on the label may correspond to a name in the literature, but do the characters of the specimen with the label correspond to characters associated with the name in the literature? Likewise, the opposite: the observations published were based on specimens which were identified. Were those specimens properly identified? Only for primary type specimens does our system of nomenclature guarantee a one to one correspondence between a name and a specimen. So, verification of names or more precisely of identifications, is an iterative process of matching names, observations and specimens, thus involving both collections and the literature. The critical points are that this data used in the process should be saved and shared and computerization is the best way to do this.

Inventory data can be captured retrospectively, but minimally it should be captured prospectively. That is, resources may not be efficiently used today to capture all data associated with insect specimens in collections except for research purposes, but the process of incorporating new material should be automated to insure that the key strokes are not wasted.

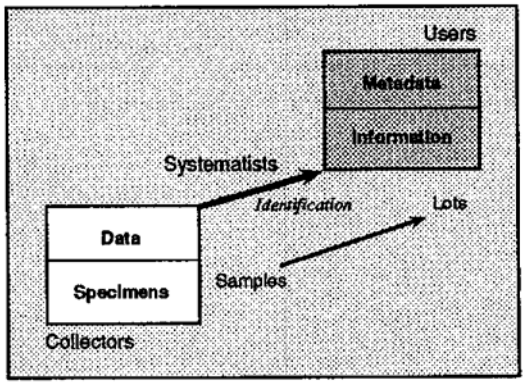
Entomological collections contain millions of specimens. To capture all the data on all the specimens seems like an impossible task. Hence, the argument goes, computerization should be restricted in scope either to higher taxonomic levels or to particular taxa or a combination of both. This argument does not, however, distinguish between the retrospective and prospective aspects. Yes, given that mass of material in collections today, retrospectively capturing all the data associated with those specimens may be an impossible task, but capturing the data associated with new incoming material is feasible and desirable, especially if automation can also aid in the preparation of that material. Likewise, computerizing data as part of research is both feasible and desirable. The key to the argument is whether the data so captured can be shared. If data are only going to be used ONCE, then how the data are handled can be determined by efficiency measures only. If, however, those data are also going to be used by others, then consideration of how the inital data capture work can be saved is desirable. Specimens all require labels for it is the label that carries the data for entomological specimens. if the process of label production is automated, then collections can prospectively build a comprehensive database of biosystematic information. Combine this with the retrospective work done as part of the research process, and entomological collections can deliver significant amounts of biosystematic information to the public.

Data Elements

Where does systematic data come from (specimens, labels, literature and people)? What does it consist of (characters, names, associations (biological, geographical and temporal) & transactions (loans & people))? And how can it be reduced to its basic elements (core fields)?

Systematic data comes from one *and only one* source: Specimens. Specimens come from one and only one source: A sample. A sample is a group of specimens collected at a single point in time and space.

Some of systematic data biological, geographic and temporal are common to all the specimens in the sample, but other data characters are particular to a specimen. Systematic data are meaningful ONLY when an identification is made. The identification process breaks the sample into lots, which therefore are only taxonomic subsets of samples. And only at the level of lots is meaningful biosystematic information available to non-systematists. The flow of systematic data into basic information is from collecting the sample, through labeling where the common data are affixed to the sample or the specimens in it, to a series of identifications which more finely subdivide the



sample making an increased number of lots of more restricted taxonomic level Order to family to genus to species to individual where ultimately all systematic data are captured and analysed to produce information.

The basic data elements of importance to systematic entomology are described below. Despite the common nature of systematic data, its limited source and flow, identification and clustering of these data elements into functionally related groups was not simple: Some data elements can be further subdivided, other could be combined, etc. Beyond this first step we have also characterized these data elements as ESSENTIAL, RECOMMENDED or OPTIONAL.

Data Structures

What are the best ways to organize these data elements into more comprehensive units (records, files and databases)? And what kinds of products (outputs) are to be derived from these structures?

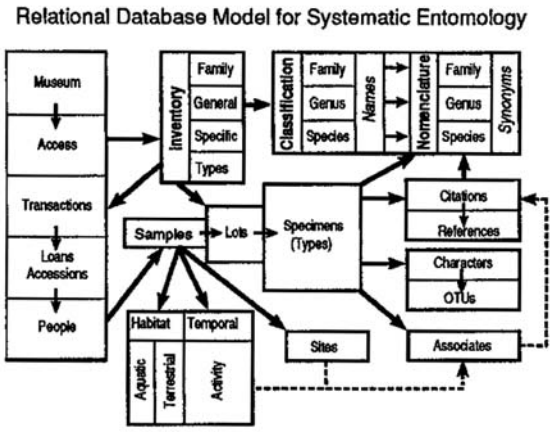
One useful structure is the relational database model. Attached is diagram of how the various elements of systematic data could be related. This model is generalized to that it could be implemented in various database management systems. Details about the model and how the various components are related are included below. From this model all the various products of systematics, from lists to monographs (see Thompson & Knutson 1987, Antenna 11: 131-134) can be derived.

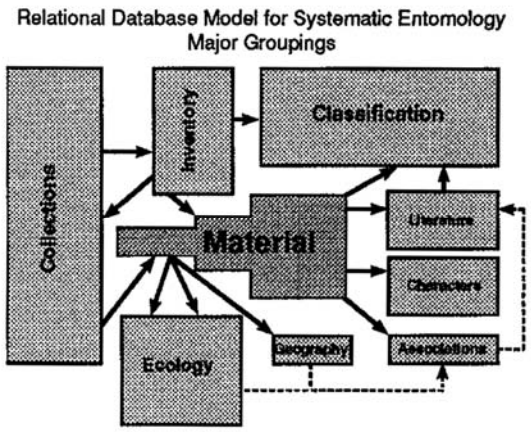
Data Standards

With standards being essential for communications and sharing of information, which of the existing standards should entomology adopt for its needs and/or what new standards need to be developed?
Numerous data standards exist for the various elements of information. And these standards range from broad and general to narrow and specific. That is, a standard may be a set of rules for creating a datum object or a standard may be a list of all the permissible variants of a datum. So, for example, we have the International Code of Zoological Nomenclature which are rules for the formulation of scientific names. Then, there is the Common names of insects & related organisms which not only includes "rules and regulations" about common names but lists all the permissible ones. Obviously, as systematists we will follow the Code, but do we want to endorse a fixed list of names for taxa? Comments and recommend-dations on various data standards are mixed in with our discussion of data elements and the data model. The only critical standard that needs endorsement at this time is a data standard for the exchange of documentation about data elements and structures. One such standard is proposed here.

Database Model and Data Dictionary

The database model and data dictionary represent a logical (conceptual) design. We attempted to identify all the critical and desirable elements of data of interest to entomological systematists. These elements were then clustered into groups, the redundant elements eliminated and relationships established between the groups. This is not a final relational database model being only in the first normal form nor a physical design which could be implemented in any particular database management system. However, systematists should critically review this view of systematic data, to see whether something is missing or not properly described, etc. A person familiar with database management systems DBMS should be able to adapt this view easily to the particularities of their software.





Two caveats are important. One, this is a complete view, but subsets can be derived from it. When subsets are derived, the user must be careful to extract all the critical data elements from related groups which are not to be included in the subset. Second, a few areas were not covered due to lack of expertise. One such area was paleontology. These areas should, however, be easily accommodated in this complete view. For paleontology, additional data elements about geological age, stratigraphy, etc., would have to be added, but these would merely form one or more new groups that would be related to either the sample or site groups.

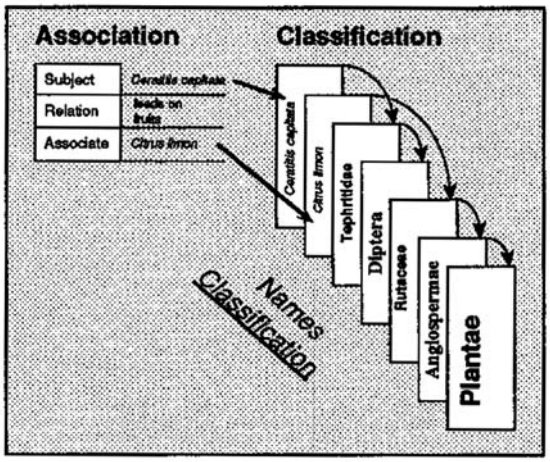
The data model is diagrammed and the data elements are listed. One diagram shows the major groups and the other shows the minor groups within those major groups. Arrows illustrate one to many relationships (Parent - Child). The table lists the data elements clustered within the major and minor groups, giving a descriptive name, short mnemonic form, the data type, the status for the element, and whether the element is used as a unique key or a link to other groups (and if so, what groups). Details about the data elements are included in the appendices. The data groups in the data model here are slightly different from those given in the appendix on collection management. A number of redundancies were eliminated from collection management when final model was prepared.

While the data model and its data dictionary should be sufficient to explain the complete view of systematic data for entomology, a few key relationships are critical to a full understanding. The core of the model is the Sample-Lots-Specimen relation, but classification is critical for the biosystematic information.

The Samples-Lots-Specimens relation reflects the flow of data into information. Specimens are collected. At that moment. there is a sample which consists of specimens (one or more) with associated data on when, where, and how these specimens were collected. That is the SAMPLE. The sample is then identified. The identification process is merely the breaking up of (or transforming, if a single specimen) the sample into taxonomic units. The level of identification may be coarse or fine, but the action is always that of assigning a taxonomic name. The taxonomic name is a unique key to the classification, which is a hierarchy of names. For insects, this process is initially done intuitively. We collect a sample of insects, then we label them, which affixes the sample data directly to the specimens, the smallest potential subunit. The specimens are then roughly identified as we distribute them to our colleagues. That distribution is actually creating the first order of lots. Eventually our colleagues finish the identification process by making a species identification. At this time, the sample data is likely to be included in a research publication or is of interest to outside users. The final lot is then actually created when the user links the sample data from the label with the species determination and creates a physical data record, If, however, in the future, the data used to create the original specimen label has been saved in a computer file and that was so indicated on the specimen label by sample number, then the user would merely have to copy the data record that the sample number identified arid combine it with the species name. The lot (a sample number plus a taxonomic name) can be further subdivided. While lots may have only a single specimen in them, the taxonomic identification level is that of a species, which is a concept for a group of individuals. So, lots, even a single specimen lots, are at least conceptually subdivisible into specimens when data are

to be captured at the level of individuals (characters). Types, especially holotypes and lectotypes, are just special specimens, or lots subdivided to the individual level Other characteristics of this relation are as the relation is transversed from Sample to Specimen: the number of data elements related increases, the number of data records (rows) increases, the amount of associated information increases, the level of identification required increases, but the taxonomic scope (level) decreases and the number of physical items (specimens) decreases. In short, all data derived from specimens are linked to some part of the Samples-Lots-Specimens relation.

Classification is merely a special hierarchical data structure, one name belongs in only one group which is itself a name. Each correct (valid) taxonomic name is always UNIQUE, so with any taxonomic name the names of all the groups to which that taxonomic name belongs can be retrieved. So, storing classification data is very economical. However, as traversing hierarchical data structures (recursion) can be difficult, so classifications and taxonomic names are usually stored in fixed structures. For example, separate data elements are used for Family, Genus and Species names. Likewise, the complete set of taxonomic names may not be available, so implementing a single hierarchical structure may not be economical. For example, for associations a complete set of taxonomic names is usually available for the subject (ex., a fruit fly), but not for the associate (ex., a plant). So given that the database would be concerned ONLY with fruit flies arid their plant hosts, the taxonomic name for the plant host could be stored in the association file. Likewise, for partial views of the data, storage of taxonomic names may be more appropriate within another file (table). For example, a small collection may want only to maintain inventory data for families. So, for such a family-level inventory system, embedding data about the family name within the inventory file may be better than maintaining separate files for classification data (see below and in appendix for more details on such an arrangement).



While the complete view of all data elements of interest is presented, most users today are interested only in partial views. Curators of small collections may only want to have inventories and/or management of data at the family level; specialists may only be interested in building catalogs (literature inventories) of their groups. All the data elements they need for their special requirements are included in the complete view. This is because we started with our special views. Hellenthal, for example, has developed systems for small collections, and I have developed catalog systems. However, those special views were combined and the redundant data elements removed to make the complete view. So, the data elements, although they are all present, may be distributed in different places. For example, for small collections that want only inventory data at the family level, some data elements from classification, nomenclature, geography must be added to the groups of main concem (Museum, Access, Inventory (family & General), Transactions, Loans/Accessions & People). Such inventory would need to have only the Biotic Region and Country (or State) from Geography, the correct family name from Classification, and all the family synonyms from Nomenclature. These data elements then would be combined with the Inventory group. For a catalog project, all the elements in classification, nomenclature, literature would be used, with some data from types, geography and perhaps associations. Again, the cataloguer would want to combine these data elements into the classification or nomenclature files. Consider the extra effort required to maintain data files, etc., necessary for the complete view just to get the location of a type. Type is a member of material, which is linked to inventory, which then is linked to collections where the subgroup museum is. So, if only the location of a type was required, then creating a data element for type location in the nomenclature file would be simpler than using the complete view.

Conclusions and Recommendations

1. Data standards are essential for efficient handling and sharing of systematic data.

2. A single comprehensive view of systematic data can support the needs of all users.

3. A relational model is the proper context in which to express this view of systematic data.

4. A common set of elements can include all useful data.

5. Given the importance of biosystematic information and its underlying sources collections and literature and the above conclusions, we recommend:

   a. that this report be endorsed as a working draft from which a final draft can be derived;

   b. that the working committee be made permanent and additional members be solicited for it as good standards must ever evolve to meet the changing community needs;

   c. that endorsement of this standards process be sought from the Systematic Resources Committee of Entomological Society of America and other interested organizations;

   d. that the initiative of the Association of Systematic Collection to develop broader standards for systematics as a whole be endorsed and the appointment of Dr. G. R. Noonan as our representative to their Task force on computerization and net working of natural history collections; and

   e. that support be sought to facilitate this standardization process and its implementation in useful programs for systematics from various funding agencies.

About this report and the working group

At the first meeting Entomological Collections Network considerable discussion was devoted to ADP standards. Hellenthal and Thompson, members of steering committee, were assigned the task of developing these standards for discussion at the next ECN meeting. This became an impossible task as distance made communicating difficult, competing priorities distract us and differing views blinded us. As the deadline approached a better way had to be found. So Systematic Entomology Laboratory provided funds to bring to Washington a few key workers knowledgeable on ADP issues and representing different viewpoints. Three days of discussion at the end of October lead to a consensus about broad issues and considerable progress on the details. The working group divided the details into three parts which were handled by different pairs. This report represents a combination of these detailed sections with the general introduction which I threw together. While I believe the introductory represents the consensus of the group, blame me for the words as the other members did not have time to review them. The sections written by others are identified with their names.