Acta zoologica cracoviensia, **46**(suppl.– Fossil Insects): 51-57, Kraków, 15 Oct., 2003

Towards a global fossil insect database

Anthony A. MITCHELL

Received: 2 Apr., 2002
Accepted for publication: 30 Sep., 2002

MITCHELL A. A. 2003. Towards a global fossil insect database. *Acta zoologica cracoviensia*, **46**(suppl.– Fossil Insects): 51-57.

Abstract. A database that will hold all the known fossil insects is presented. Database design is discused and the progress towards collecting data is reported.

Key words: fossil, insect, computer, database, holotype

Anthony A. MITCHELL, 10 Prinys Drive Wigmore Gillingham Kent ME8 ORB UK; Maidstone Museum and Art Gallery, St Faiths Street Maidstone Kent England. ME4 1LH
E-mail: ed@ mbcmus1.demon.co.uk

I. INTRODUCTION

Following his contribution to BENTON's "Fossil record 2" (1993), Ed JARZEMBOWSKI realised that only a continual update of information could make it possible to easily produce a subsequent edition. Getting information from CARPENTER's (1992) "Treatise" on fossil insects had shown that the "Treatise" was by no means complete. Edna CLIFFORD, as honorary abstractor, volunteered to produce a card index file on new taxa of fossil insects by extracting data from papers supplied initially by Andrew ROSS and later by Ed JARZEMBOWSKI and other workers. These were given to her in batches, a year at a time from 1982 onwards, the cut off date for the "Treatise". Her brief was to search for 'nov. gen'. or 'n. sp.' and to fill in a card with the Author, Date, Title, Publisher, Specific name, Family and Order. The cards were then sorted under Author in year blocks. Eventually with almost 2000 cards, the manual system became almost impossible to use and a computerised system became imperative.

II. COMPUTER DATABASES

There are two main kinds of computerised database. The simplest is a s p r e a d s h e e t. All the data is held in a grid, with each horizontal line holding all the data about one species in headed columns such as Name, Author, Title, etc (Table I). The whole database can be reordered alphabetically on any column and searched for any key word. Some columns may remain empty or contain the same information many times. A disadvantage of the spreadsheet design is that every piece of information must be entered every time. This could include, for example, the title of a single paper covering dozens of species. Any slight spelling mistake, especially in a key word, could result in an apparent loss of data. More sophisticated spreadsheets can look up and copy previously entered data from the same column. If a change needs to be made in any of the data, this could be very time consuming, as every item must be checked independently.

Table I

Example of part of a Spreadsheet Database (title truncated to fit the page)

Name	Author	Title
Abaristophora nepalensis	DISNEY & ROSS 1996	Abaristophora & Puliciphora (Diptera, Phoridae) from Dominic
Abaristophora domicamberae	DISNEY & ROSS 1996	Abaristophora & Puliciphora (Diptera, Phoridae) from Dominic
Aberrokorynetes abludens	WINKLER 1990	Two new genera of fossil Korynetinae from Baltic Amber (Coleoptera)
Aboilus femineus	GOROCHOV 1996	New Mesozoic insects of the superfamily Hagloidea (Orthoptera)
Aboilus krassilovi	ZHERICHIN 1985	Jurassic Insects of Siberia and Mongolia: Orthoptera
Aboilus pullus	GOROCHOV 1996	New Mesozoic insects of the superfamily Hagloidea (Orthoptera)
Aboilus tigris	GOROCHOV 1996	New Mesozoic insects of the superfamily Hagloidea (Orthoptera)
Aboilus zebra	GOROCHOV 1996	New Mesozoic insects of the superfamily Hagloidea (Orthoptera)
Accretonemoura grata	SINITCHENKOVA 1987	Historical development of stoneflies (Plecoptera)
Acixiites costalis	HAMILTON 1990	Insects from the Santana Formation, Lower Cretaceous of Brazil
Acixiites immodesta	HAMILTON 1990	Insects from the Santana Formation, Lower Cretaceous of Brazil

A r e l a t i o n a l d a t a b a s e is a series of linked spreadsheets called tables. Each item in a table is numbered and linked to a corresponding number in other tables. One table could hold Name data while another holds Author and Publication data (Table II). This greatly reduces the amount of data that needs to be typed in, and, as it appears only once, editing and correction of errors is much easier. The programme takes care of linking the numbers between the fields 'Author ID' in the two tables.

The relational database chosen for computerising the cards was Microsoft Access. This is part of the Microsoft Office suite, which allows easy conversion to Microsoft Word and Microsoft Excel and was already installed on the Maidstone Museum computer. Microsoft Office is readily available worldwide. Access is easily capable of holding all the data expected to be amassed. An advantage of a Microsoft Access-based database is the ability to automate much of the data entry and therefore save time and, at the same time, check for consistency. For example it has been set to prevent the duplication of entries in certain fields.

III. EDNA

The computerised database (called EDNA after Edna CLIFFORD), was originally designed simply to hold the data recorded on the card index file. Its limited purpose was to provide the data required for a "Fossil Record 3". This publication would only require information at family level, but Edna was extracting down to species level. As computer work progressed, it was found that, to save space and time, she had sometimes shortened titles. Fortunately only one change was required in the 'Title', field of the 'Tref' table, the relational part of the database taking care of the rest (Table IV). A bigger problem was that she had sometimes omitted the family name and superfamily or subfamily given instead. This is easily spotted, as the suffixes are different. Suborder, infraorder, division and order were more difficult to unravel without consulting the original paper, especially as some authors had moved higher taxa within the Linnean hierarchy, sometimes with no explanation. EDNA now contained the fields subfamily, family, superfamily, group (the informal level 'group' has been included to accommodate the various divisions between superfamily and suborder), suborder, order, author, title, publication, volume, date and page number.

As the database grew, it started to become a practical taxonomic supplement to CARPENTER's "Treatise".

Table II

Example of the same data as a Relational Database. The ID numbers would be assigned automatically. When linked through the ID numbers, the result will be the same as for the spreadsheet

Name	Author ID
Abaristophora nepalensis	1
Abaristophora domicamberae	1
Aberrokorynetes abludens	2
Aboilus femineus	3
Aboilus krassilovi	4
Aboilus pullus	3
Aboilus tigris	3
Aboilus zebra	3
Accretonemoura grata	5
Acixiites costalis	6
Acixiites immodesta	6

Author ID	Author	Title
1	DISNEY & ROSS 1996	Abaristophora & Puliciphora (Diptera, Phoridae) from Dominic
2	WINKLER 1990	Two new genera of fossil Korynetinae from Baltic Amber (Coleoptera)
3	GOROCHOV 1996	New Mesozoic insects of the superfamily Hagloidea (Orthoptera)
4	ZHERICHIN 1985	Jurassic Insects of Siberia and Mongolia: Orthoptera
5	SINITCHENKOVA 1987	Historical development of stoneflies (Plecoptera)
6	HAMILTON 1990	Insects from the Santana Formation, Lower Cretaceous of Brazil

IV. ESF

The ESF meeting of the fossil insects network at Dijon, 1997 produced a 'wish-list' of information that a specimen based fossil insect database should ideally contain (Table III). Having already experienced the difficulty of extracting even simple taxonomic data from publications in several languages, it seemed unlikely that such a complex database could ever be produced. To be of value, data must be complete and reliable. A figure of a million specimens was mentioned at Dijon. To simply enter a million items, once the data had been found and verified, would take a minimum of 2 x 1,000,000 minutes = 4000 working days. Doing corrections could increase this time by an order of magnitude. It was suggested that a simpler way would be to merge museum records. These are often in purpose-made databases that are often incompatible with each other but could be merged, in theory at least, if sufficient computer programming time could be hired. As each museum records different data, some more than others, the 'wish-list' would still be far from complete. Validation is a far bigger problem. When new specimens were added to museum collections, their identification will have depended on the knowledge of the identifier and the literature available at the time. Since 1985 at least 3,000 new species have been named of the estimated 40,000 total. As many museum collections go back to the 19th century, and often have not been revised since they were accessioned, they will have been given names erected before that date. Unless they are absolutely identical in all respects to the holotype, it is possible that the identification is wrong and must not get into the database. When old collections, and even relatively new ones, are looked at in detail it is often apparent

Table III

"Wish-list" for a specimen based database. For further explanation see HIRSCH-MEYER 1997. = indicates a link with another part of the database
* indicates that this information is included in EDNA

Specimen
Number
Taxonomy=
Locality=
Horizon
Site
Collection=
Place
Date
Collector
Kind
Part
Sex
Growth stage
Taphonomy=
Status
Description
Picture
Author=
Comments

Taphonomy
Preservation
Articulation
Orientation
Comments

Taxonomy
Phylum
Class
Order *
Suborder *
Superfamily *
Family *
Subfamily *
Tribe
Genus *
Subgenus *
Species *
Subspecies
Author *
Year *
Description
Collection=
Authority
Comments

Stratigraphy
Absolute age
Era *
Period *
Subperiod *
Superstage
Stage *
Substage
Source / Author
Comments
Series *
Group *
Formation *
Member *
Bed *
Source / Author
Comments
Zone
Subzone
Horizon
Source/Author
Comments

Localities
Coordinates
Sedimentology=
Other taxa
Stratigraphy=
References=
Picture
Sites *
Map
Comments

Geography
Palaeogeography
Geography

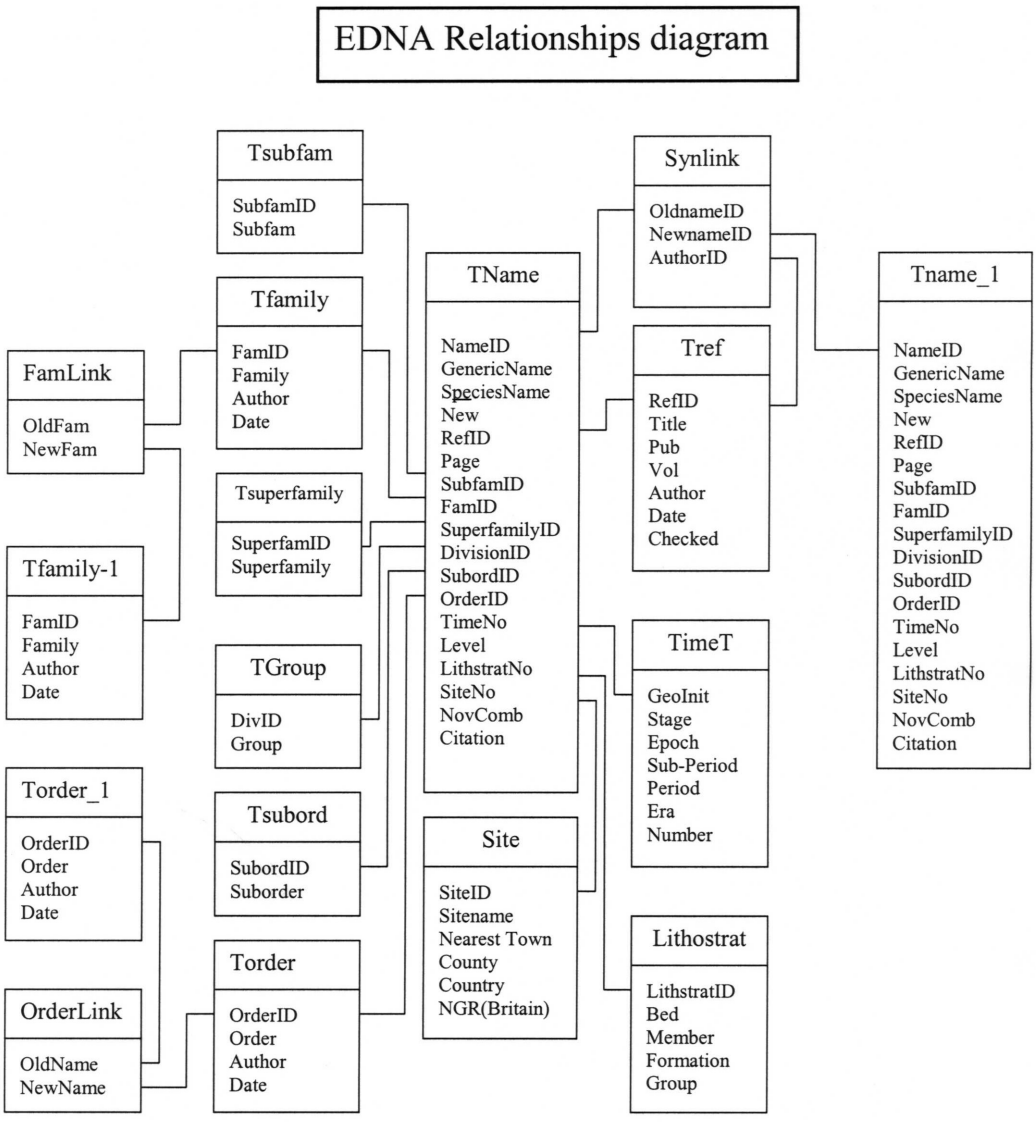
Collections
Address
Person in charge
Facilities
Former collection
Taxa present=
Numbers=
Comments

Bibliography
Author(s) *
Year *
Original title
English title *
Source
Kind of source
Pages *
Figures
Comments

Environment
Biofacies
Fauna
Flora
Lithofacies
Rock type *
Diagenic minerals
Sedimentary structures
Interpretation
Climate
Source / Author
Literature=

Table IV

EDNA relationships diagram. A box is called a table with the table name at the top. Each item in the table is called a field.
ID = identification number, used to link data between tables
Torder_1, Tfamily_1 and Tname_1 are copies of Torder, Tfamily and Tname produced by the program when required. These tables contain both valid and obsolete names. The link table is used to find the valid name when an obsolete name is entered or vice versa. Number in TimeT allows the geostratographical column to be displayed in chronological order



that the geographical (site) and lithological data is vague or even incorrect. Every item must therefore be scrutinised before entry, which would be impossibly time consuming.

The easiest part of the Dijon 'wish-list' to add to the EDNA database was site, lithology and geological time. Time would also be necessary for "Fossil Record 3" and any "Treatise" update. At first the data was found simply from the title, then by rereading all the source publications. At the same time, other taxonomic details were added. In non-English publications it was sometimes impossible to find the non-taxonomic data, and even in English they were sometimes hidden in the text and require a lot of finding, if they were recorded at all. As only holotypes have been included in EDNA, only the type location is recorded.

V. GENERAL PROBLEMS

Three main problems have been encountered whilst extracting data directly from the literature.

1. L a n g u a g e. Computers are very pedantic over spelling and cannot find words spelled even slightly differently. For example, key in 'Brasil' and the computer will fail to find 'Brazil'. Key in 'Espana' and it won't find 'España', 'Espagne' or 'Spain'. Where an English title has been provided, it has been used in preference to the original language. Russian and Chinese scripts have been transliterated or translated. As there are several different transliteration conventions from Chinese and Russian into English, the same word could appear in different spellings. This is a special problem with site names and authors so one spelling has been adopted over another when it is highly probable that the names are referring to the same place, and cross-references have been made where necessary. All accents have been ignored on the assumption that all keyboards have non-accented keys and key words will be less likely to be missed if none are accented. An exception is the titles, which do have the correct accents as it is unlikely that workers will want search a title for a key word. Typing in accented letters using a standard English keyboard is quite difficult. For example to get 'ö' requires holding down the Alt key and typing the code 0246. Workers without the character map codes might find it difficult to type in 'España' An alternative would be to duplicate the title in the original language but this would need a field width greater than 250 characters, which would not fit on a line of A4, even in landscape format. All words that have had the accents removed have kept to the original spelling so that 'ö' becomes 'o' and not 'oe'.

2. T a x o n o m y. There is a tendency for workers in one specialised field to upgrade superfamilies to suborders, suborders to orders and to introduce more levels of hierarchy between family and class. The informal level 'group' has been used to accommodate some of these extra levels where they may be useful, otherwise a 'traditional' and stable taxonomy has been adopted based largely on the "Treatise". The data entry form is designed to look up previous entries at family level so that a family cannot appear in two higher groupings at the same level. In general, where there is a conflict, the most recent classification is taken as correct unless the author seems out of step with the majority and gives no reason for the systematic change.

3. S y n o n y m y. In the light of research, species often change generic names and sometimes family or even higher taxa. It is important that any such changes are reflected in the database and only the most recent name is used. At the same time the older names must remain available. To accommodate invalid names, EDNA already included some pre 1983 species as synonyms.

VI. CONVENTIONS IN EDNA

All names that have since been superseded have an = sign after them. It is then an easy matter to look up the most recent name using the built-in query facility. Working the other way, the database will also look up all the synonyms for any valid name. As a quick way of including all the genera in each family, all the generic names have been entered from the "Treatise". As specific names,

authors, site and time details are sometimes not available from that source, the symbol $ has been used after the generic name. When the genus and species is encountered in the primary literature, and the extra data found, these records will be ammended.

D i a c r i t i c m a r k s. Article 11.2 of the International Code of Zoological Nomenclature says that a name when first published must "have been spelled only in the 26 letters of the Latin alphabet" but that deviation from this rule does not invalidate the name. Article 27 also states that "No diacritic or other mark …. is to be used in a scientific name". When an author has coined a new name as "*Aus*" *bus* or *Aus*? *bus* the quote marks and query have been removed i.e. open nomenclature has been ignored. It follows that the = and $ signs mentioned above are not to be taken as part of the name. Subgenera are included in brackets.

VII. DATA EXTRACION

There is almost no limit to the combinations of data that can be displayed. Complicated or simple searches can be made. If the database becomes available on CD and Access 2000 is loaded, queries can be customised to include counts and graphs.

These are a few that are built in (they should occupy one line when printed on A4 paper).

1. Search for a particular taxon and display all species recorded for that taxon. Searching at family level will give all recorded sub families, genera and species. At order level, suborders, families, subfamilies, genera and species are displayed.
2. Search for a specific taxon and display species and author details.
3. Search for a specific taxon and display species site and time data.
4. Search for a specific age or site and display species and other taxonomic details.
5. Search for author and display publications.
6. Search for author and display species.
7. Find synonyms, either the valid name for an invalid name, or all included names for a valid name.

On July 4th 2002 EDNA contained 7150 species, 3900 genera, and 1295 families from references (including synonyms).

T h e F u t u r e

At present, EDNA is running in Access 97, but it is hoped that it will be updated to Access 2000. This will make it possible to offer more of the database on the World Wide Web than can be currently found in *Meganeura*. All the taxonomic data from the "Treatise" has now been entered. This will not include every species, site or time data until all primary sources have been consulted, but will have every genus. The eventual aim is to include wing venation diagrams and publish the whole database on CD to be run on any computer containing Access 2000 or later versions. The present database is strictly holotype based but with a very small modification can be used to store records of any species from any site. This facility could easily be built into the CD.

REFERENCES

BENTON M. J. 1993 (ed.) The fossil record 2, 845 pp., Chapman and Hall.

CARPENTER F. M. 1992 Superclass Hexapoda. Treatise on Invertebrate Paleontology, Part R, Arthropoda 4, 3 & 4: 1-655.

HORNSCHEMEYER T. 1997 The fossil insect database. *Meganeura*, **1**: 15-16.

International commission on zoological nomenclature. 1999. International Code of Zoological Nomenclature. Fourth Edition. XXIX + 306 pp. International Trust for Zoological Nomenclature, London.

JARZEMBOWSKI E. A. 1995 Palaeontology's own 'Blue Book'. *Inclusion,* **19**: 16.