Technical report | 70

Towards a global biological information infrastructure

Challenges, opportunities, synergies, and the role of entomology

Edited by:
H. Saarenmaa and E. S. Nielsen †

Project manager:
Hannu Saarenmaa
European Environment Agency

European Environment Agency

European Environment Agency
Kongens Nytorv 6
DK-1050 Copenhagen K
Tel. (45) 33 36 71 00
Fax (45) 33 36 71 99
E-mail: eea@eea.eu.int
Internet: http://www.eea.eu.int

Contents

Preface

Towards a global biological information infrastructure: Challenges, opportunities, synergies, and the role of entomology

We have now enetered the 21st century. The world is going towards Information Society. For entomologists this time is particularly challenging because of the wealth of data that is potentially available in this field. Being able to share data efficiently would allow entomologists to make a major contribution to the conservation of biodiversity. The combination of new technologies with systematics and collections based research may offer an opportunity to strengthen such activities in the future. There are many good ways of framing the activities such as the Clearing House Mechanism (CHM) and Global Biodiversity Information Facility (GBIF).

This all calls for a new approach. Biodiversity informatics and taxonomy are emerging as information sciences. We believe that if we are able to create a useful information infrastructure for entomology, it should directly address the burning questions of the time, such as the slow rate of discovery of new species, and extinction that follows from lack of knowledge and value on biodiversity. If data, information, and knowledge could be shared more efficiently than has been the case in the past, it would increase the credibility of the taxonomic community in the eyes of funding organisations, and have a positive snowball effect over a wide range of activities.

The papers in this volume are results of a one-day symposium that was held during the XXI International Congress of Entomology in Iguassu Falls, Brasil, on 24 August 2000. The symposium was called upon to make an inventory of the ongoing activities and possibly to lay down some foundations for further cooperation among the various projects. Twelve presentations were made. Seven of them were turned into papers during the Autumn of 2000 and are printed in this volume. Four other papers that covered 1) Entomology at the Costa Rican InBio, 2) Beetles and beetle larvae of the world: An interactive identification and information systems for families and subfamilies, 3) Developing and sharing data globally: The Global Butterfly Information System GLOBIS, 4) The BioSystematic Database of World Diptera: the first global master species database, are available as abstracts in the Congress volumes. There also is a website that links to all the presented systems (1).

Looking at the list of projects and the systems presented, it all looks very exiting. Yet the bigger picture might be still missing. Is there interoperability between the systems? If we compare entomological information management with other areas, such as plant information, it is easy to realise that we still have some way to go. How these challenges will be met was covered by Ebbe Nielsen in the opening speech on the GBIF.

En route to the first meeting of the GBIF Governing Board, the co-editor of this volume Ebbe Nielsen passed away on 7 March 2001. The worldwide entomology and biodiversity informatics communities sustained a huge loss. We dedicate this small work to his memory.

April 2001
Hannu Saarenmaa

(1) http://www.eionet.eu.int/Topic_Areas/Nature_Protection_Biodiversity/Biodiversity/GBII

The Tree of Life project

A multi-authored, distributed Internet project containing information about phylogeny and biodiversity

David R. Maddison ([2]), Wayne P. Maddison ([3]), Jeremy Frumkin ([4]), and Katja-Sabine Schulz ([2])

Abstract

The Tree of Life project (ToL) is a collaborative effort among biologists to portray the relationships and characteristics of organisms. Experts on groups of organisms synthesize available information and portray their view of the phylogeny of that group, including discussion of evidence and alternative hypotheses, alongside additional information about the organisms' characteristics. The ToL is currently a series of static HTML web pages, but in the near future it will be converted into a dynamic, database-driven system. Presentations of the information in the ToL database will then be customizable, allowing the project to better serve a diversity of audiences. The ToL database will be able to communicate with other databases, serving phylogenetic and other information about a group of organisms to other databases, and in turn receiving additional information about taxa from other databases.

Keywords: evolutionary tree; organismal characteristics; database.

The affinities of all the beings of the same class have sometimes been represented by a great tree… As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.

(Darwin, 1859)

Introduction

Organisms we see today are but leaves on the tips of Darwin's Tree of Life. The diversity of species arose by the branching of the evolutionary tree, and the diversity in form of these species by evolutionary change along those branches. As the evolutionary tree is the conduit along which the genes (and therefore traits) of organisms flowed, it is not surprising that knowledge of the shape of this phylogeny can be critical for understanding modern biodiversity (e.g., Ridley, 1983; Felsenstein, 1985; Harvey and Pagel, 1991; Maddison and Maddison, 1992; Martins, 1996; Pagel, 1999).

The Tree of Life Project (http://phylogeny.arizona.edu/tree/phylogeny.html) uses phylogeny as the central organizing principle for information about organisms and biodiversity. It is a collaborative effort among biologists providing a collection of information, available over the Internet, about the phylogeny and diversity of life on Earth. It consists of a series of web pages, each illustrating and discussing an individual species or a group of species, linked together in the form of a current view of the evolutionary tree of life. Along with pictures and introductory information of interest to the general public and students of all levels, Tree of Life pages feature specialized sections (on morphology, phylogeny, biogeography, etc.) addressing the needs of researchers in the field. There are currently over 300 biologists in 21 countries authoring pages of the Tree.

The Tree of Life Project (ToL) currently has three primary goals: (1) to provide comprehensive and authoritative information on the phylogenetic relationships among all species of organism, living and extinct (a goal that will never be fully achieved); (2) to provide information about the characteristics of groups of organisms; (3) to provide information on every species of organism.

(2) Department of Entomology, University of Arizona, Tucson, AZ, 85721, USA, tree@ag.arizona.edu
(3) Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA
(4) University of Arizona Library, University of Arizona, Tucson, AZ, 85721, USA

An information infrastructure for German insect collections including multimedia and GIS tools

Karl-Heinz Lampe ([8]) and Klaus Riede ([8])

Abstract

The German Ministry of Science and Education has launched the EDIS-project (Entomological Data and Information System) to digitise and harmonise the rich, but scattered entomological collections housed at various German institutions. The concept is illustrated for the DORSA-subproject, which will integrate German Orthoptera collections within one 'Virtual Museum', accessible by an internet-based user interface. DORSA is a network project, connecting expertise in data-basing, collection management, systematics, geographical information systems and neuroinformatics. The core of DORSA is a specimen-based database of important grasshoppers and crickets in German collections. The taxonomic backbone will be the 'Orthoptera Species File', a global species register already available on the World Wide Web. DORSA integrates specimen-based pictures and sound recordings. The species-specific songs will be used as a knowledge base for the development of song recognition algorithms and bio-acoustic 'Rapid assessment tools'. In addition, all localities will be geo-referenced, resulting in a huge data-set of point data, which can be intersected with other GIS-maps (e.g. on rainforest distribution). A customised Java-tool allows geographic depiction and retrieval of taxonomic data.

Keywords: Orthoptera, species database, specimen database, collection management, GIS, song recognition

Introduction

The number of insect species on earth and their actual extinction rates is a matter of speculation for several years already, but exact numbers are still not available (Stork et al. 1997). This is mainly due to the lack of an efficient information infrastructure. Complete registers of valid taxa only exist for few insect groups, and much of the information stored within museum collections is not readily accessible, because most of it is not digitised. In many institutions it is common notion that 'computerisation' of collections might be possible for the vertebrate departments, but that the mission will be impossible for invertebrate sections due to the overwhelming number of species and specimens, compared to the lack of staff and money. Further difficulties for insect collection managers are the high number of undetermined specimens or undescribed (new) species ('taxonomic impediment').

Nevertheless, an impressive demonstration of feasibility has recently been accomplished by the Insect@thon project (Komen & Marais 2000), where around 21 000 insect inventory records of the Namibian National Museum have been entered by 92 schoolkids on 1 weekend. The project managers stress the need for digital access to the hand-written catalogues of huge first-world collections such as the Natural History Museum London, with 65 million insect specimens: 'We estimate that some 70% of these collections originate in the third world. Inasmuch, we strongly believe that first-world museums are urgently accountable to us…' (http://www.natmus.cul.na/biodive/insectresults.html). Other initiatives such as CONABIO or InBio show that biodiversity-information management in developing countries is much more advanced than in many developed countries. Especially European institutions seem to have severe difficulties in adopting the new information technologies.

(8) Zoologisches Forschungsinstitut und Museum Alexander Koenig (ZFMK), Adenauerallee 150-164, D-51113 Bonn, Germany, E-Mail: k.lampe.zfmk@uni-bonn.de; k.riede.zfmk@uni-bonn.de.

This asymmetry was part of the rationale for the establishment of the Global Biodiversity Information Facility (GBIF) (Edwards et al. 2000). As part of this initiative, the German Ministry of Science and Education (BMBF) has launched the EDIS-project (Entomological Data Information System) to digitise and harmonise the rich, but scattered entomological collections housed at various German institutions within one specimen-based collection database (http://www.insects-online.de/).

The EDIS-project consists of separate and self-dependent subprojects including 'global species registers' (cf. GLOBIS, this volume) and collection databases of specimens with a connection to geographical information systems (GIS). Further projects are rapid assessment tools for automatic identification at the molecular level, by optical analysis of bee's wing venation and sound recognition of crickets and grasshoppers (http://www.insects-online.de). The respective databases will be pooled by an Oracle-based database, which will provide Internet-access (SYSTAX: see http://www.biologie.uni-ulm.de/systax).

This paper deals with the subproject 'German Orthoptera collections' (DORSA, according to the German acronym). Germany harbours rich collections including material from tropical countries dating back to the 19th century (Figure 1). Most of this material is not data-based. The DORSA project will digitise specimen-specific information for crickets and grasshoppers and integrate them within one 'Virtual Museum', accessible by an internet-based user interface.

Figure 1. **German research collections with estimated numbers of Orthoptera holotypes. Most of the material is not data-based**

Databasing and collection management

In practice the overall efficiency of data-basing the inventory of traditional entomological collections depends on two factors: suitable software, and management measures to ensure the highest possible data quality already during the input process. All data entry is based on determined specimens (systematic label information), but 'determination' is not limited to the species level - it can be any higher taxon (e.g. 'Acrididae' for an unknown grasshopper). In our institution we introduced a lock-step programme for ergonomic and efficient data entry consisting of the following steps:

1. primary data capture of systematic label information
2. validity check of the systematic information against a current catalogue
3. set up of a collection based catalogue of taxa (consisting of the updated systematic information including synonyms, hierarchy of taxa, authors, year, etc.)

4. secondary data capture of sampling information (such as locations, collectors, determinators etc.)
5. validity check of the geographic information against a current gazetteer (by adding geographical attributes such as latitude, longitude and in a hierarchy such as province/ area, state, country, continent/ocean and a link to the zoogeographical region)
6. set up of a collection based catalogue & completion of lists (e.g. collectors, determinators)
7. final data entry of existing specimens into the database

Each step of the procedure is clearly separated from the next. Therefore, everyone involved in this work can clearly see his own area of responsibility as well as the progress of his work. A nice side effect is the allocation of the various jobs involved where they are most welcome. Someone who is interested in working with catalogues by looking for further geographical or systematic information can work easily together with someone who is more interested in doing an accurate data capture. That means a single person is no longer forced to complete all the various tasks alone. Yet another advantage is that any of these procedures can be stopped or interrupted and even taken over by a third party with very little extra effort.

Systematic backbone

The 'Orthoptera Species File' (OSF) will be used as a taxonomic backbone. The OSF is an electronic catalogue of named grasshoppers and crickets, including pictures and sounds, and is available on the WorldWide Web (OSF: Otte and Naskrecki 2000). The OSF is one of the few fully functional global species registers. Queries allow searching for valid names as well as synonyms, their taxonomic reference and the depository place of the holotype. This means that one can ask for all holotypes for a certain museum, as known from the type descriptions. DORSA will realise the next step: a link between species names and existing specimens in German collections. In the future, a simple mouse click on a taxon should produce a list of specimens, together with a map of point data.

Geographical Information System (GIS)

All data sets refer to a locality. To connect them to a GIS, they have to be geo-referenced by their geographic coordinates. In the ideal case, they have been determined exactly, for example by a 'Global Positioning System' (GPS). In most cases, localities are given as geographic names, and coordinates have to be determined afterwards by searching gazetteers or atlasses. In many cases, locality information is vague, which requires coding of imprecision. Geo-referencing is a time-consuming process, but can be speeded up during the project by building up a thesaurus of specific collection sites and major collector's routes.

Once geo-referenced coordinates can easily be exported into a GIS and plotted as a distribution map or analysed by geo-processing. For example, the intersection with borders of states or provinces produces calculated species lists for administrative units. These lists are useful for conservationists and decision makers, but their maintenance is a time-consuming task. There are numerous additional applications for biodiversity maps in GIS-format, among them: Comparison of maps from different sources and different projections; Calculation of biodiversity hotspots; Intersection with other GIS-layers such as eco-regions, land use, population pressure or climate change predictions, to name just a few.

Figure 2 shows an example for the potential of GIS analysis of collection data. Point data based on 3.578 data-sets of a ZFMK Homoptera collection (with 7.969 specimens) are plotted on a satellite view of the world (Figure 2). The original data for each specimen can be requested by simple mouse-click at the locality point.

Figure 2. **World map with localities of the ZFMK Homoptera collection, superimposed on a satellite picture**

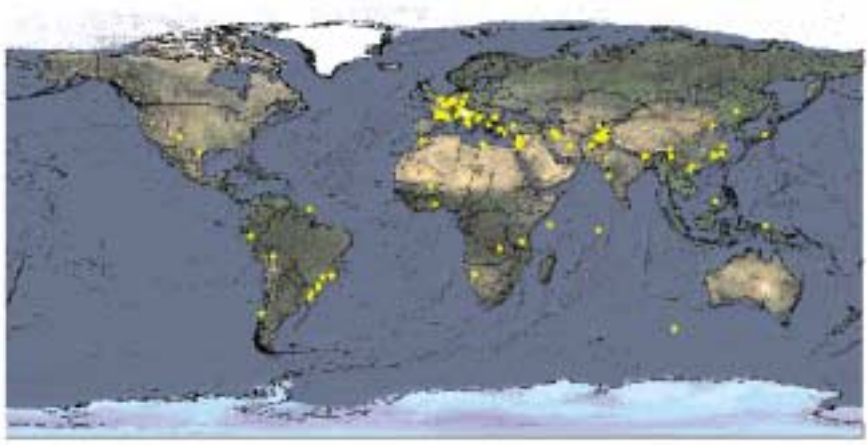
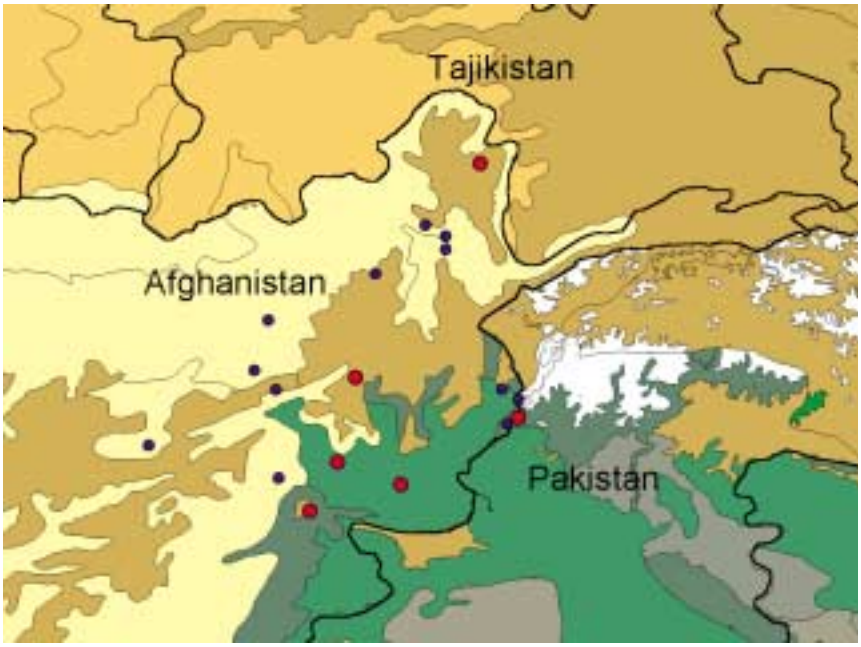


Figure 3. **Ecoregions of Central Asia and collection sites (red dots: origin of type material)**



With a GIS one can easily zoom into the world map and enlarge special areas and/or change the background information. Figure 3 shows parts of Afghanistan and Pakistan structured by ecological attributes such as the vegetation type. Dots represents collection sites. GIS analysis can be greatly enhanced by pooling data sets from different collections (see Spatial Analyst, this volume).

In spite of these advantages, biologists are still reluctant to use GIS tools. One of the reasons is the user-unfriendlness, such as lack of a standardised query language. The introduction of desktop GIS improved this slightly, but there are still many problems. Even simple questions such as: 'How many species occur within a certain area?' require a number of complex operations. Therefore, a new Graphical User Interface (GUI) has been developed to publish interactive maps on the World Wide Web in cooperation with the Geography Department, Bonn (Fitzke & Friebe 1999). It consists of a platform independent, Java- based information desk for a combined display of geographic and database information (see a pilot version used by the 'Global Register of migratory species': http://www.groms.de). The information desk will be adapted for DORSA to allow queries such as: Show distribution map of a given species! Show maps of non-described species within a higher systematic taxon such as a family! Show

material collected during the Kaiserin-Augusta-Fluss-Expedition (Papua-New Guinea 1920). Or even strange sounding queries such as: Show all species singing with a carrier frequency lower than 2 kHz! Results will be depicted as point data maps.

Multimedia and neuroinformatics

A special feature of crickets and grasshoppers are their species-specific songs which can be used for rapid species identification (cf. Riede 1993). Germany has a long tradition in Orthopteran bio-acoustics and neuroethology, which resulted in several important sound archives ('phonotheks') at universities and in private hands (for reviews on Orthopteran communication see Ragge 1998, Riede 1998). At present, the majority of these data is stored on analog media such as magnetic tapes, film or videos. These recordings are now digitised and stored in a standardised format (e.g. wav-files for acoustic data). File names together with data on the original analog data source (tape number, deposit, etc.), recordist and localities are entered into DORSA. Voucher specimens exist for some, but not all of the recordings. Sound files could be used either to analyse inter- and intraspecific song variation, or to provide input for automatic song recognition algorithms. Neural networks for song classification and identification at the species level are presently developed in close cooperation with the Neuroinformatics Department at Ulm University.

The aim is a bioacoustic 'rapid assessment tool' for non-invasive mapping and identification of Orthoptera in the field.

Perspectives

DORSA will be accessible by Internet from any part of the world as one 'Virtual Museum Collection', which is important for potential users in species-rich, but resource-poor developing countries with incipient biodiversity infrastructure. The 'Virtual Museum Collection' will help to improve classical taxonomic work such as description of new taxa. Further important functions of this database such as distribution map generation and retrieval of pictures and songs from determined specimens, will be especially useful for ecologists, conservationists and applied entomologists.

At present, types and paratypes are entered into DORSA. The process of databasing is an excellent opportunity for type revision, lectotype designations and eventually repatriation of secondary types. In the case of type loss, re-collection of topotypes should be initiated. Topotypes should also be collected and designated for the country representing the 'terra typica', in particular for endemic species. The database will reveal information about historic species distributions which can be compared with actual distributions. Especially in rainforests, such a comparison will form the base for estimates of the actual conservation status and insect extinction rates, which at present are not even informed guesses.

Acknowledgements

DORSA forms part of the Entomological Data Information System (EDIS) and is funded by the German Ministry of Science and Education (BMBF). We thank Sigfrid Ingrisch (ZFMK Bonn, Germany) for helpful discussions about cricket taxonomy, Daniel Otte and Piotr Naskrecki (OSF: Academy of Natural Sciences of Philadelphia) for kind supply with Orthoptera species authority files, and Christian Dietrich (Neuroinformatics department, Ulm University, Germany) for providing preliminary results of his Ph.D. thesis.

References

Edwards, J.L., Lane, M.A. & Nielsen, E.S. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289, 2312–2314.

Fitzke, J. und Müller, M.(2000): Simple Features in der Praxis:

OpenGIS-Strukturen in Auskunftssystemen für Umwelt- und Naturschutz. In: Cremers, A.B. und Greve,K.(Eds.): Umweltinformation für Planung, Politik und Öffentlichkeit (Environmental Information for Planning, Politics and the Public)., Beiträge zum 14. Internationalen Symposium 'Informatik für den Umweltschutz', 'Umweltinformatik aktuell', Band 26, Marburg 2000, pp. 321–329.

Komen, J., Marais, E. 2000. The Insect@thon project. http://www.natmus.cul.na/biodive/insectresults.html

Otte, D., Naskrecki, P. 2000. Orthoptera Species Online. http://viceroy.eeb.uconn.edu/Orthoptera.

Ragge, D.R., Reynolds, W.J. 1998. The songs of the grasshoppers and crickets of Western Europe, Harley Books in association with the Natural History Museum, London.
Riede, K. 1993. Monitoring biodiversity: Analysis of Amazonian rainforest sounds. Ambio 22, 546–548.

Riede, K. 1998. Acoustic monitoring of Orthoptera and its potential for conservation. Journal of Insect Conservation 2, 217–223.

Stork, N. E., Adis, J., Didham, R. K. (Eds.) 1997. Canopy arthropods. London: Chapman & Hall.