# THE LEPIDOPTERA TAXOME PROJECT

http://www.ucl.ac.uk/taxome/

Expression of Interest for
Integrated Project under
EU Framework VI
Call identifier: EOI.FP6.2002

*presented on behalf of*
Muséum National d'Histoire Naturelle, Paris
National Museum of Natural History, Leiden
The Natural History Museum, London
State Museum of Natural History, Stuttgart
Stockholm University
Tor Vergata University, Rome
University of Amsterdam
University College London
University of Copenhagen
Verlag für interaktive Medien Gbr
*…and others*

James Mallet
Honorary Research Fellow, The Natural History Museum, London
&
Professor of Biological Diversity
University College London
4 Stephenson Way
LONDON NW1 2HE, UK
and
tel: (+44)-20-7679-7412
http://abacus.gene.ucl.ac.uk/jim/

DRAFT: 23 May 2002

**The Taxome Project**

## 1) Need and relevance

***Background.*** Taxonomy, the naming and classification of organisms, is fundamental to all biological science and biotechnology. Taxonomy is the original bioinformatics: its modern incarnation dates from the scientific revolution in 17th Century Europe, and was formalized using Linnaeus' 18th Century system of binominal nomenclature and classification.

Taxonomy today receives meagre funding in even the richest countries of Europe, while in the USA, funding is increasing. Perhaps its long history encourages the view in Europe that little remains to be done. Yet, compared with other major informatics projects such as geographic mapping, meteorology or genomics, our knowledge of taxonomy is woefully incomplete. Estimates of the actual numbers of species on this planet vary over two orders of magnitude (3-100 million)[1], and species relationships remain obscure in even the best-known groups. Taxonomic information is scattered widely in small-circulation books and print journals, and users of taxonomy have no simple way to access data even for the ~ 1.5-2 million "known" species[1] so far described.

The inadequacy of existing taxonomy informatics has triggered repeated calls from major scientific and public figures[1,2] for modernization and increased funding. For example, Sir Robert May, now Lord May, President of the Royal Society of London and formerly Prime Minister's Science Advisor in the UK, has frequently pleaded for global taxonomy databases in public speeches over the last ten years[1]. Prof. C. Godfray is only the most recent supporter of a global web-based presentation of taxonomic knowledge[2], a concept we here term **"the Taxome Project"** (i.e. complete taxonomy for all of life). Yet taxonomy in Europe continues to decline, and the virtual lack of training in Western European countries is becoming critical[3,4]. Taxonomy is a casualty of changes in fashion, coupled with a normally productive emphasis on competitive research in many European countries: fashion ensures virtually no grant funding or overheads are available for taxonomy in Europe, so that competitive and cash-strapped universities and museums are loth to commit personnel to the taxonomic enterprise.

A major innovative program for taxonomy on a European scale is proposed here to galvanise research towards the global Taxome Project. This funding would stem the decline in taxonomy and stimulate a revival across Europe as the major institutions compete for funds for innovative taxonomic research projects.

***Why Europe needs the Taxome Project.*** Taxonomy is the basis upon which all other biological technologies are built. However, many immediate applications of large scale taxonomy justify urgent funding to strengthen and modernize taxonomy across Europe:

*a*) To provide expertise for ***pharmaceutical and biotechnology industries***, which contribute strongly to European exports. Taxonomy and taxonomic expertise are required as a basis for bioprospecting and gene discovery[4];
*b*) To ***identify organisms that threaten human health and vectors of pathogens***.
*c*) European agriculture depends on ***identification and monitoring of European pests and diseases*** and their vectors in agriculture and animal health, as well as of potential novel crops and food sources;

c) To enhance utility of ***genome databases***. Gene sequences are carefully validated for sequence content. However, identity of the organism(s) that give rise to the genetic data is not checked, making much genome information potentially useless;

*d*) To ***obtain basic understanding of species, biodiversity and evolution***. ***Biogeography*** and the emerging field ***macroecology***, which investigate how biodiversity is distributed geographically[5,6], crucially depend on accurate taxonomy, for instance in understanding how Europe was recolonized since the last ice age[7];

*e*) To ***monitor human-induced influences on biodiversity***, for example due to ***global climate change***[8,9];

*f*) To ameliorate the ***"taxonomic impediment" preventing effective assessment of global biodiversity*** in Europe and partner developing countries[10]. European governments have ratified the ***Convention on Biological Diversity*** and are required to monitor their own biodiversity, as well as to provide expertise and repatriate data to aid poorer nations to do the same.

The Taxome Project outlined here falls under EU Framework VI theme 1.1.6.3 "Global Change and Ecosystems". It will "contribute to the sustainable use of land ... resources", and aid in "assessing and forecasting changes in biodiversity structure ..." within the "Biodiversity and Ecosystems" subtheme. The research should also contribute to related themes including 1.1.2.1 "Applied Information Society Technology", 1.1.2.4 "Knowledge and Interface Technologies", and 1.1.7.1 "Knowledge-based Society and Social Cohesion".

***State of the art in Europe, and need for a European effort.*** European major museums and systematic institutes hold ***over 80% of the types*** of the world's biodiversity, amassed mainly during Europe's colonial past ("types" are individual voucher specimens on which names of taxa, such as species or genera, are based)[11]. The Natural History Museum in London alone is thought to possess over 60% the world's type specimens of butterflies and moths (Lepidoptera)[11]. These historical collections mean that a European consortium will form the most suitable base for development of a modern taxonomic infrastructure.

As well as holding most of the primary data, Europe ***has already initiated important taxonomy bioinformatics projects*** and our expertise in taxonomy and informatics is still comparable to that in the Americas, in spite of generally poorer funding and recent declines in the number of taxonomists. In the USA, taxonomy is currently experiencing a renaissance, with new and strengthened NSF support for systematics, taxonomic inventory work, and especially the new NSF-funded program *Assembling the Tree of Life*[12]. Given pre-existing data and expertise in our major systematics institutes, Europe will be well-placed to capitalize from the Taxome Project, and will help redress the current North Atlantic taxonomy imbalance.

**2) Scale of ambition and critical mass**

***The Taxome Project.*** The global Taxome Project will ultimately collate taxonomic information for all organisms, both living and fossil, into a single public, online database[2]. A variety of current European[13] and world[14] initiatives already aim to supply world-wide taxonomic and/or specimen-based information. None of these global projects has reached an advanced stage of completion.

Taxonomy since its inception has been vital to biology because of adherence to a set of nomenclatural rules, today codified by international commissions of biological

nomenclature. Taxonomic nomenclature and classification has no formal legal status in any country (as far as we are aware), yet it has nonetheless enabled information about taxa such as *Solanum tuberosum*, *Escherichia coli* and *Drosophila melanogaster* to be communicated internationally with little ambiguity. The value of these voluntary taxonomic standards is incalculable, and any project such as this must work within these international guidelines even while modernizing taxonomy. The international consensus which has allowed biological knowledge to flourish is similar to, but arguably more important and certainly longer-lasting than the value of standards such as HTTP and HTML in providing access to information on the internet.

The revolutionary approach to taxonomy advocated here will require ***global cooperation between experts for each group of organisms***. Needs and solutions will differ markedly between taxonomic groups (for instance between bacteria, plants, insects and birds). The entire Taxome Project is consequently highly desirable, but an unwieldy goal for any one research consortium of reasonable size. Major inroads will instead be made by tackling large but coherent sections of the overall tree of life.

**3) Objectives and proposed achievements of the Lepidoptera Taxome Project**.

Our consortium is therefore argues for a taxon-oriented approach, using ***The Lepidoptera Taxome Project*** both as a demonstration or model workpackage and also as a practical and useful solution for a megadiverse group of insects (Lepidoptera, or butterflies and moths). In addition, our knowledge of Lepidoptera demonstrates that an important subset, ***The Butterfly Taxome Project,*** can be completed to a very high level of detail, while the overall Lepidoptera project can reach a very useful online nomenclatural catalogue level within 5 years. Taxonomic knowledge of actual organisms is most crucial and considerably more demanding than the necessary computing and DNA-based technologies, which can be readily moulded to any project. The project depends on and exploits existing high standards of taxonomy within our consortium.

***The Overall Objective*** (and expected % completion) ***of the Butterfly Taxome Project*** will be ***to develop and produce public online databases containing major taxonomic information for very large numbers of species***:

*a*) ***Complete taxonomic catalogue*** consisting of classification and lists of all known species-group and higher level nomenclature including synonymy for the world's 17,500 species (99.5%).
*b*) ***Complete taxonomic literature reference database*** for all valid names and synonyms (99.5%).
*c*) ***Digitized original descriptions,*** type specimen data and collection information (99%).
*d*) ***Visual information***, such as photographs of types and drawings of anatomy (99%).
*e*) ***Key new revisions*** in poorly known groups, particularly within the Hesperiidae, Lycaenidae, and Nymphalidae: Satyrinae (long-term: 60% of worst known groups).
*f*) ***New Molecular data for phylogeny estimation*** will require a new European "horizontal genome program" (sequencing the same few genes in many species) as well as world-wide coordination (e.g. Oregon State University, Harvard University, Stanford University, Milwaukee Public Museum, University of Alberta, working under NSF's Tree of Life program[12]). Several key gene sequences are emerging as standards in Lepidoptera phylogeny estimation: mitochondrial COI/COII, nuclear *EF-1a*, *wingless*, and some other nuclear genes, possibly *6-phosphogluconate dehydrogenase* and others.

Overall, around 2Kb of mtDNA and 4Kb of nuclear DNA unique sequence are required for the phylogenetic work envisaged. Data will be deposited in GENBANK[15] and linked via our online taxonomy (98% of genera).

*g*) ***Classification and "aligned" phylogenetic data***. Phylogeny estimation requires aligned character data so that comparable characters (including DNA nucleotides) are unambiguous. Morphological character matrices and aligned sequence data will be made available on the Taxome Project Website, or via links to TREEBASE[16] (95%). The classification used in the online database will depend on estimates of phylogeny using molecular and additional morphological characters.

*h*) ***Life-history information***, host plants; diagrams and pictures of young stages (50%).

*i*) ***Basic distribution data***, especially of type localities (95%); as well as links to distribution mapping data, for example in European distribution databases or via specimen databases[13].

*j*) ***Custom-built taxonomy database software*** to present this information reliably and flexibly. End-users will consist of specialized taxonomists and other biodiversity experts, conservation managers, and society at large. With many conflicting potential users, advanced database software is needed to allow flexible extraction of information at varying levels of detail. Some excellent examples of web-based software demonstrating some of the principles required are already operational[17].

The much larger ***Lepidoptera Taxome Project*** as a whole will achieve for all Lepidoptera parts 3(*a*-*c*) of the ***Butterfly Taxome Project***, as outlined above, during the 5 year funding period, and to approximately the same degree of completion. For the first time, this work will enable accurate counts of the numbers of described species and large-scale surveys of the areas of Lepidoptera systematics that need further work.

The Lepidoptera Taxome Project is valuable to European strategic aims because:

*a*) ***It will provide a model for taxome*** collation and delivery to end users.

*b*) ***The Lepidoptera (butterflies and moths) contain a substantial fraction of all described species***. The Lepidoptera form ~ 10% of all described species of animals, plants and microbes living today[1,18]. Butterflies (Lepidoptera: Rhopalocera) consist of 17,500 species, ~ 1% of the world's described species. This project will make significant inroads to the ***Taxome Project*** as a whole.

*c*) ***Many Lepidoptera are important pests of agriculture*** for instance the European corn borer (*Ostrinia nubilialis*), corn earworms and budworms (*Heliothis* and *Helicoverpa* spp.)., and the *Brassica*-feeding *Pieris* and related genera. Taxonomic knowledge of pests in their natural environment is essential for their control, particularly via environmentally sustainable techniques such as biocontrol and genetic methods.

*d*) ***Butterflies are key indicators for monitoring ecosystem health, biodiversity and conservation***[19]

*e*) Taxonomic knowledge will enable monitoring of the ***biotic impact of global climate change***, as already demonstrated using European butterflies[8,9].

*f*) ***Extensive training of postgraduate students and postdoctoral fellows*** in taxonomy and molecular systematics will enhance European competitiveness in biotechnology and biodiversity studies for the future.

*g*) Because butterflies and moths are easily recognizable, or even iconic to the European public, the work will communicate the value of biological research to Europeans of all ages.

**4) Integrated activities and resources required**

***Overall.*** The research outlined above requires extensive integration of research at major systematics institutions holding type specimens and important modern collections, and of molecular geneticists and systematists, both within Europe and elsewhere. Novel computer science is required in the form of software and database development, and social science research is needed to maximize accessibility of taxonomy to the widest possible array of end-users in Europe and the rest of the world.

***Taxonomy and systematics research.*** In additional to salaried staff already employed by the institutions in the consortium, the Lepidoptera Taxome Project aims to train 20 new postdoctoral fellows and 20 PhD students, and similar numbers of technicians distributed among the consortium, over a period of 5 years. Half of these will work directly with the taxonomy, that is in curation, type identification and recording, nomenclatural issues, morphological work, revisions, catalogue construction, and biological information-gathering (life-histories, distribution data, etc.). The rest will work on database and interface technologies for online information delivery, DNA sequencing and phylogeny reconstruction, and research at the interface of this work and conservation, global change, and other end users. Adequate laboratory equipment (microscopes etc.), computing resources, high-throughput DNA sequencing facilities, and enhanced specimen curation are also required. Type specimens and also signficant Lepidoptera systematics expertise are also found in Eastern Europe (Russia, Poland, Hungary, Czech republic, among others) as well as in the West. It is vital that EU-associated states be incorporated into this effort.

***Management structure.*** Management of the entire project will be based initially in London (UCL). A major taxonomist will be hired to manage the taxonomic core of the project (possibly Dr. Gerardo Lamas-Müller, who, with international collaborators in GloBIS has masterminded the construction of a Tropical America synonymic catalogue, still unpublished, consisting of 7600 species). Finally, the projects' activities require lower level secretarial and technical support, together with an effective overall project management team.

***Intercontinental collaboration.*** Integration of the European project must be achieved with other experts world-wide. For instance, the Butterfly Taxome Project is a result of pioneering work by GloBIS[20], consisting of scientists from USA and Peru, as well as 3 European groups included within our consortium. Integration is also required between the Lepidoptera Taxome Project and other leading taxome and online specimen-databasing projects[13,14] in order to contribute effectively to the Global Taxome Project of all taxa.

**Annex 1. Cited literature**


1 Gaston KJ, May RM 1992 Nature 356: 281_282
2 Godfray HCJ 2002 Challenges for taxonomy. Nature 417: 17_19
3 Disney RHL 1998 Nature 394: 120
4 Jaspars M 1998 Nature 394: 413
5 Gaston KJ, Blackburn TM 2000 Pattern and Process in Macroecology
6 Bell G 2001 Science 293: 2413_2418
7 Hewitt,G 2000 Nature 405: 907_913
8 Parmesan C et al. 1999 Nature 399: 579_583
9 Thomas CD et al. 2001 Nature 411: 577_581
10 Hoagland KE 1996 http://www.ascoll.org/Newsletter/taxImp.htm
11 Estimates based on existing data collated for Riodinidae, Nymphalidae and Geometridae held at the NHM, London. The lack of information makes even these mild claims hard to assess; the Taxome Projects outlined here will, among other achievements, enable an assessment of the global value of European and other important world type collections for the first time.
12 NSF 2002 http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf02074
13 For example: http://www.cetaf.org/, http://www.biocase.org/, http://www.insects_online.de/, http://www.faunaeur.org/enbi/info.html, http://www.sp2000.org/, http://viadocs.essex.ac.uk/html/
14 For example: http://www.all_species.org/, http://www.gbif.org/, http://www.ento.csiro.au/globis/, http://www.sis.agr.gc.ca/pls/itisca/taxaget, http://tolweb.org/tree/phylogeny.html
15 http://www.ncbi.nlm.nih.gov/
16 http://www.treebase.org/treebase/
17 For example http://www.insects_online.de/, http://www.nrm.se/ve/pisces/acara/welcome.shtml, http://viceroy.eeb.uconn.edu/Orthoptera
18 Wilson EO 1992 The Diversity of Life
19 http://www.butterfly_conservation.org/
20 http://www.ento.csiro.au/globis/


Annex 2. Organization of the consortium

Institutions and hierarchical layout of the entire Taxome Project

To be completed!