Nota lepid. 22 (1): 35–47; 01.III.1999 ISSN 0342–7536

HOSTS: a database of the host plants of the world's Lepidoptera

Gaden S. Robinson

Department of Entomology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

Summary. The Natural History Museum's HOSTS database is intended to provide eventually a thorough inventory of the host plants of the world's Lepidoptera. The methods used for data capture, the editing and validation processes, the database structure and the inherent limitations of the project are described. The current status of the database, its actual and potential products, and possible directions for future development are outlined, and the problems in making it widely available while safeguarding intellectual property rights are discussed.

Zusammenfassung. Die Datenbank HOSTS am Natural History Museum, London, hat eine umfassende Zusammenstellung der Wirtspflanzennachweise für alle Lepidopterenarten der Erde zum Ziel. Methoden der Datenerfassung, der Herausgabe- und Validierungsprozess, die Struktur der Datenbank und die systembedingten Grenzen des Projektes werden beschrieben. Ferner werden der gegenwärtige Stand der Datenbank, ihre aktuellen und potentiellen Nutzanwendungen sowie mögliche Richtungen für die künftige Weiterentwicklung vorgestellt. Ein wichtiges Problem besteht darin, Urheberrechte an geistigem Eigentum mit den Anforderungen an eine weite Verbreitung der gespeicherten Information in Einklang zu bringen.

Résumé. La banque de données HOSTS du 'Natural History Museum' à Londres a comme objectif de fournir un inventaire exhaustif des plantes-hôtes des Lépidoptères du monde. Les méthodes employées pour la collecte des données, pour l'édition et les procédés de validation, pour la structure de la banque de données et les limitations inhérentes du projet sont décrits. L'état actuel de la banque de données, les produits actuels et potentiels qu'elle livre et les directions pour un développement futur sont également décrits, ainsi que les problèmes posés par sa mise à la disposition générale en ce qui concerne la protection des droits de propriété intellectuelle.

Key words: Lepidoptera, host plants, world resources, inventory database, data processing, intellectual property rights.

Introduction

Information on what eats what and where in the complex web of relationships between caterpillars and plants is of use to a very wide range of users. The demand for such information is

increasing as a corollary of the increase in demand for information (and rapid access to it) by, notably, environmental and agricultural interests. The provision of such information requires access to data that has both geographical and taxonomic breadth. With some 135,000 recognised Lepidoptera species feeding potentially on more than a quarter-million species of plants, the eventual size of an even remotely credible databank is considerable.

A great deal of information on Lepidoptera host plants is already available for, notably, Europe and North America. But most is in either printed (i.e., published) or manuscript form, the latter often as card indexes, the former scattered in an enormous literature that covers three centuries. Attempts have been made to compile regional compendiums of host plant data. These include, for example, that by Tietz (1972) for the Macrolepidoptera of North America, and by Emmet (1992) for Great Britain and Ireland. The only attempt at a global compendium is that by Zhang (1994) for Lepidoptera of economic importance.

The integration of Lepidoptera host plant data and the efficient sorting, indexing and interrogation of that data to answer a wide range of questions for a wide range of users necessitates its being in the form of an electronic database. In the late 1980's staff of the NHM Lepidoptera Section began collecting data in electronic form as a series of pilot projects. In 1993 we developed the concept of a large database that would provide eventually world-wide coverage and be flexible enough to deliver both printed and electronic products to a user base that included amateur entomologists and professional biologists involved in systematics, conservation, agriculture, forestry, biocontrol and quarantine regulation. This concept has evolved into the HOSTS database.

Comparatively small and specialised data sets are used routinely by systematists to enrich the data content of taxonomic treatments by providing an ecological context. The observed host plant ranges of small groups of Lepidoptera may well, by their uniformity, reinforce hypotheses of relationship: larvae of *Utetheisa* (Arctiidae), for example, feed on Leguminosae and Boraginaceae, and morphologically distinct species-groups are restricted in their feeding to one or other host family.

Easily accessible data on Lepidoptera host plants allows the conservationist to at least predict the presence of particular insects

in a habitat given the requisite botanical information. It also permits the setting of clear diversity objectives in habitat enrichment and restoration — a "wish-list" of Lepidoptera species can be matched against a list of the host plants necessary for the species to establish themselves.

Rapid access to data on the insects attacking particular plant species (rather then vice versa) is vital to applied entomologists. The recent discovery of novel damage to cypress foliage in nurseries in East Anglia required a rapid response and using the HOSTS database we were able to narrow the possibilities to two probable and two possible North American Argyresthiidae species in less than five minutes. Eliminating three of the four (leaving one of the two "probables") by checking voucher specimens against a near-comprehensive reference collection took another ten minutes. This identification would have taken considerably longer using conventional means.

Access to comprehensive or near-comprehensive host plant data is also invaluable in biocontrol studies, helping to suggest appropriate groups and regions for further investigation, to narrow searches and to warn of potential problems in species that are not host-specific.

HOSTS, while at present by no means giving universal coverage, provides us with the wherewithal to interrogate a large databank to find the host plants or host plant ranges of a species, or group of species, and to do the opposite and search for the larvae that feed on a plant or group of plants. We can perform these searches at all taxonomic levels and limit searches by country or zoogeographic region. We can examine the numerical structure and frequency distribution of host plant utilization (Robinson, 1998), examine correlations, and provide printed compendiums and indexes listing either plants and the larvae that eat them, or larvae and their host plants at geographical scales from country to global.

In this paper the methods used for data capture, database structure, editing and validation processes and the inherent limitations of the project are described. The current status of the HOSTS database, its actual and potential products, and possible directions for future development are outlined, and the problems inherent in making it widely available while safeguarding intellectual property rights to the entire compilation are discussed.

Data acquisition and data sources

Our early priorities were to abstract major printed compendiums of host plant information together with major manuscript resources to provide a large and credible base of information that could then be further developed by the addition of sources that contained fewer records but were complementary to the major sources. Examples of major printed sources that were abstracted include McGugan *et al.* (1958–1965), Tietz (1972) and Scott (1986) — North America, Silva *et al.* (1968) — Brazil, Yunus & Ho (1980) — Malaysia, Emmet (1992) — Great Britain and Ireland.

Major manuscript resources included: Edward Meyrick's ledger of the host plants of the world's Microlepidoptera, culled from correspondence, literature and the many thousands of specimens that passed through his hands in the space of some sixty years from about 1876 to 1936; the card index compiled by Comstock and Henne for North American Microlepidoptera that complements Tietz (1972) and which is housed in the Los Angeles County Museum of Natural History, and the comprehensive card catalogue of Nymphalidae host plants developed by Phillip Ackery (The Natural History Museum (NHM), London).

Progressively smaller sources were included as the project progressed; literature searches and the polling of fellow-specialists for suggestions of key works for inclusion resulted in a steady accumulation of data. In 1995 it was decided that North America would be the first geographical priority for full development of the database, followed by the Oriental region.

Electronic and manuscript lists of host plants were solicited from colleagues both in the NHM and elsewhere. A demonstration database was established on the World Wide Web and information solicited either as e-mail, word-processor files, databases or directly via a WWW input form (see below). The response to this "public appeal" was surprisingly generous; large data sets donated include Japanese Lepidoptera on Fagaceae (Dr. N. Teramoto), world Lycaenidae (Dr. Konrad Fiedler) and California butterflies (Ms. Marian Fricano). We would be delighted to receive additional contributions!

Abstracting was carried out, for the most part, by volunteers (work experience students) and by students undertaking vacation

work. The minimum usable information — names of Lepidoptera and host plants — was typed into a temporary Paradox database together with any additional relevant information such as abbreviated details of damage, locality and cited (secondary) sources. The field structure of the temporary database was restricted to the bare minimum required and expanded only later to the complete format (see below). Repeated information, such as author and date of the source, was added subsequently as a global change. The accuracy of transfer was, overall, surprisingly good and the tenacity and responsibility of our work-experience students, some as young as fourteen and dealing with a subject entirely novel to them, was laudable.

Despite initial optimism, comparatively few sources proved suitable for data acquisition by optical character recognition — the narrative rather than tabular form of most sources precluded efficient conversion and even where the format was suitable, poor print quality often resulted in an unacceptably high level of error in conversion. But OCR has proved useful in some cases and we consider it a valuable adjunct to abstracting by manual methods. In these cases the source is scanned using a Hewlett-Packard Scanjet 6100C and OCR performed to deliver ASCII text using Omni-Page (Caere). From this, column-tabulated files are generated (for checking) then converted into delimited ASCII text using WordPerfect 6.1 and imported into Paradox.

Database structure

HOSTS is a "flat" database comprising 23 alphanumeric fields totalling 313 characters. Abstractors fill a maximum of 15 of these fields (* — asterisked), but the abstracting of a single source typically involves only seven or eight fields. Unfilled fields are either left blank, filled globally, filled from other relational databases, or involve check "signatures" as part of subsequent editing and validation. The fields are as follows (field length is in brackets):

Family (5): Abbreviation of family-group name, e.g. NOCTU(idae), derived from first five letters; ambiguities such as HELIOdinidae and HELIOzelidae are resolved by adaptation, e.g., HELID and HELIZ. This field and the next are entered automatically by relational linking to a database of the generic names of the Lepidoptera (derived from Nye (1975–91) and developed by B. R. Pitkin) and act as a check on spelling of the generic name.

Subgp (5): Abbreviation of (usually) subfamily derived as above.

NCA (3): Name Check Authority — an entry indicates verification that the insect name is currently valid and comprises initials of checker or source used (e.g., CLE indicates that the name used is compatible with *Checklist of the Lepidoptera of Europe* by Karsholt & Razowski (1996)).

Genus (20)*: Insect generic name.

Species (20)*: Insect species name.

Subspecies (20)*: Insect subspecies name.

Author (16)*: Insect author(s) — in full unless exceeding field length; names are abbreviated according to a table of standards (e.g., Hübner, Denis & Schiff.).

Damage (20)*: Succinct damage descriptor which may be abbreviated (e.g., in leaves, on fls/fruits/leaves); "in" is used specifically to denote internal feeding or specified concealed feeding ("in rolled leaves"); the use of "on" tends to be somewhat generalised in the literature but we have tried to restrict its use to external or unspecified concealed (but not internal) feeding. This field may also include indications of, for example, ant associations ("on flowers + ants").

Plantgenus (17)*: Genus of host plant or, in the occasional case of a carnivorous larva, the insect prey. Very rarely plant and insect genera have identical names and ambiguity is avoided by suffixing an insect generic name with a "Z".

Plantspecies (20)*: Species of host plant or prey insect.

Plantsubspecies/var (20)*: Subspecies or varietal name of host plant.

Plantfamily (17)*: Family of host plant. This field is entered automatically by relational linking to a database of the generic names of plants (derived from Brummitt (1992)) which acts as a check on spelling and current validity of the generic name. The terms *polyphagous* and *detritophagous* may be used in this field with appropriate modifiers in the *Damage* field. Other non-standard terms used are: *Algae; Filicopsida* (i.e., unidentified fern(s)); *Fungi; Insecta* (i.e. predaceous — with generic name of host in the *Plantgenus* field (see above); *Lichenes; Musci.*

PCA (3): Plant check authority. As NCA above. "JTK", for example, indicates the name is valid in Kartesz's (1994) checklist of the vascular flora of North America.

Locality (20)*: Country from which the host record originates. Large countries (e.g., Brazil, USA) are subdivided into states and the state entered from a table of standard abbreviations (e.g., USA: TX). Occasional captive rearing records (see below) refer to rearing of stock originating from one country on a "substitute" food plant in another. Provenance data for species involved in such "hobby rearings" is recorded as, for example, "Ecuador (prov.)". A relational database can be used to provide a link from this field to the major zoogeographic region from which the record originates.

Source (16)*: The source from which the record was abstracted or received (i.e., the primary source). This may be an author's name (e.g., "Fletcher", "Brown *et al.*"),

indicate a manuscript or database source, ("Meyrick MS"; "Intachat db") or an unpublished source that cannot be consulted ("Jones pers. comm."; "WWW input"). Source details are held in a bibliography maintained as a word-processor file; this will be converted eventually into a database.

Date (6)*: Date of the source; field size permits use of square brackets where source date is determined by external evidence.

Secondary source (16)*: If the source cites an earlier publication as the original source of the record this is entered here; a blank field may not guarantee that the source is the original. Abstractors have often had difficulty in identifying secondary source citations; in the abstracting of some sources, secondary citations were ignored.

Sec-date (6)*: Date of secondary source.

Original name (30): Entered globally as a concatenation of *Genus* + *Species* + *Subspecies* fields (above) immediately after abstracting; permits back-tracking of the name in the original reference. This was not recorded in the early stages of database development and an entry followed by "[R]" indicates retrospective entry. Retrospective entries may not match the version of the name used in the source.

Original host name (30): Entered for the host plant as the preceding field, and with the same limitations.

CR (1)*: Captive record: the plant is not known to be a host in the field but is accepted by larvae in captivity. The entire record is only included if development was completed successfully. Entries are "Y" (yes), "N" or blank (no), or "?" (maybe).

Reliab (1)*: Reliability. Possible entries are "?" (record dubious), "N" (record is an error), or "O" (oviposition — only — observed). Doubtful identifications of insect or plant are indicated by suffix queries in the name fields (e.g., "*Solanum tuberosum* (?)" [*Solanum*, but only maybe *tuberosum*]; "*Solanum* (?)" [maybe a *Solanum*]; the latter is vague enough to earn also a "?" in the *Reliability* field.

Nathost (1): Possible entry is "N" when the host plant is definitely not native to the country or area where the record originates. Many crop plants and ornamentals, for example, are not native to the countries in which they are grown. This field is not used in this phase of database development.

Editing and validation of nomenclature

Once abstracting is complete, new records are reformatted with the full field structure of the HOSTS database and any global fills (such as source and date) are performed. The *Original name* and *Original host name* fields are filled as described above. Subsequent phases of editing and validation involve cycles of progressive refinement, to correct mis-spellings and to modernise and standardise insect and plant nomenclature. Problem entries

are carried over into another cycle. Once a substantial number of records have been prepared for editing, the automatic checking of the insect and plant generic names described above provides an opportunity to correct spellings and plant generic synonymy. Author names are standardised and plant and insect names are then checked electronically against two databases of previously-validated names. These databases allow us also to correct automatically many frequently-encountered plant and insect synonymies and to convert many common vernacular names of plants. Names that still fail to achieve a validation check are then processed against the other nomenclatural databases available to us (such as Missouri Botanical Gardens' database of Peruvian plants and Scoble's (NHM) catalogue of Geometridae) and recombined or synonymized as necessary to achieve a contemporary and consistent nomenclature.

Remnant non-validated names are then checked manually against recent authoritative sources (checklists and monographs, the Missouri Botanical Gardens VAST database on the WWW) and, as a final resort the NHM Lepidoptera systematic card catalogues (most families updated only to 1982) and *Index Kewensis* on CD-ROM (which only records the existence of a plant name but does not give its current status).

Editing and rendering current and consistent the insect and plant nomenclature and other elements in the HOSTS database is the most time-consuming part of the operation. It inevitably throws up inconsistencies between regional taxonomies for plants and Lepidoptera and requires compromise.

Limitations, accuracy and problems

No global catalogue of Lepidoptera host plants can ever be comprehensive. Neither can its content ever be entirely accurate. While it would in theory be possible to search and abstract the entire world's entomological literature, the resource implications of such a task are monumental. So in practice a strategy of prioritisation is needed to achieve a credible compilation using finite resources. The strategy adopted here is that of abstracting the largest sources first then adding complementary key works, as described above. In a few instances, complementary key works may include the entire oeuvre of a particular author who has

specialised in the monographing of reared material (e.g., Kumata on Gracillariidae).

While it would be admirable to trace all host plant records to their original source, this is not considered practical for the entire Lepidoptera. However, in developing adjunct databases to HOSTS for the world's bombycoid moths and the Neotropical Rhopalocera respectively, Kitching & Beccaloni (in prep.) are attempting just this. Several authors (e.g., Sattler, 1967 for Palaearctic Ethmiidae) have published catalogues of host plants for smaller groups in which all records have been back-tracked to their original source.

The potential sources of error in any compilation of host plant records are manifold. At the root lies misidentification of either plant or insect by the original observer or recorder. Further errors may occur in the transcription of records, a classic case being the confusion of *E.*[*ugenia*] *malaccensis* with *E.*[*ndospermum*] *malaccense*, which resulted in persistent citation of Myrtaceae as a host family for Uraniidae rather than Euphorbiaceae (Lees, pers. comm.). Confusion may occur between similar or identical plant and insect names (e.g., *Aristotelia* — Gelechiidae or Elaeocarpaceae) or a transcriber may confuse similar generic names such as *Asperugo* (Boraginaceae) and *Asperula* (Rubiaceae). Confusion may occur when the context is in a language with which the abstractor or transcriber is unfamiliar. A history involving synonymy that is later reversed may result in a perfectly valid host record being switched from one species to another.

Rearing caterpillars obtained from eggs from a captured female on a host plant found acceptable by trial and error may result in the publication of a host record that is erroneous in that the relationship is entirely artificial. Such laboratory rearings are not always clearly cited as such.

Erroneous host plant records are cumulative — repeated citation gives them a spurious authority and they are extremely difficult to detect and delete. As errors accumulate, there is a danger that the "noise" of different erroneous records may obliterate a correct insect-plant relationship especially if this is a single observation on a unusual host plant.

Much of the original abstracting for this work was carried out by volunteers unfamiliar with plant and insect nomenclature

and unsure of the meaning of some contexts. Manuscript sources were not always perfectly legible. So there is potential for further errors being added in the abstracting process. Resources did not permit us to trace all records to origin nor to check all abstracting work.

HOSTS: current status

At the time of preparation of this paper (June 1998) the HOSTS database contained 102,981 records covering:

	Lepidoptera	Plants
Families	99	299
Genera	5315	3103
Species	20,457	10,216

Geographical coverage of the database shows some regional bias, reflecting geographical priorities for abstracting. Numbers of records for each major zoogeographic region are:

Regions	Records
Palaearctic	20,088
Nearctic	38,793
Neotropical	8295
Afrotropical	6774
Oriental	20,623
Australasia	4384

There are additionally some 251 records from Hawaii, 164 from New Zealand and 3609 records that are attributed either to the Holarctic region or have no location attributed. "Pantropical" and "cosmopolitan" species are counted here as if they were from the Oriental region.

HOSTS: the products

It is intended that a series of major products from HOSTS should include printed compendiums of data for at least some of the major zoogeographic regions covering all Lepidoptera and plant groups. Such a compilation has just been completed for North America (Robinson *et al.*, in prep.) and one is in preparation for the Oriental region. Poorly served by existing published sources, a host plant catalogue for the Afrotropical

region is also badly needed but development of the database to the point where this can be produced will necessitate additional funding.

Other medium-term products envisaged include catalogues of the Lepidoptera that feed on particular plant families. The current level of interest in legume biology, for example, suggests that this would be an appropriate group for a global catalogue.

The HOSTS database can also be used to generate data for question-driven research and collaboration in this area is currently being solicited, and an extension of preliminary work on frequency distribution of host plant specificity is envisaged.

HOSTS on the WWW and the future

The current demonstration database on the Web
http://www.nhm.ac.uk/entomology/hostplants
contains some 3000 records and is intended to publicise the HOSTS project as well as provide useful public-domain information. The database search program permits the user to search by genus or species name of plant or insect and then to perform cross-referencing searches. The site also includes an input form that allows the user to contribute individual records to the database and includes a request for additional information in a variety of electronic formats (see above).

The value of HOSTS as a look-up tool for specific Lepidoptera-plant associations is inestimable, and it is inevitable that, in response to demand, the entire database will be made available for search on the Internet in the near future. However, the resources that have been devoted to the development of HOSTS are such that Internet availability cannot jeopardise the published products and other possible commercial or academic applications of the database.

The search structure of the present WWW database would result in an unmanageably large number of records being returned to the enquirer if the database were to be enlarged and would permit downloading of unacceptably large slices of the dataset if it were applied to HOSTS. We envisage that HOSTS will be mounted on the Web with a search routine that requires the user to narrow his search by taxonomic (plant and insect) and geographic criteria, and possibly omit fields until the number of

records retrieved falls below a specified limit. Only then will data be transmitted to the enquirer.

Further into the future, we envisage rapid growth of metadata handling capabilities in which cross-linking to other databases will be possible. This will allow the almost instantaneous retrieval of supplementary taxonomic data on the plants and the insects, geographic information, illustrations and, indeed, all that we can presently retrieve by pulling the drawers of reference collections and combing the shelves of libraries together with much, much more.

Pooling resources of host plant information is just one way in which we can propel data into the public domain. We would be delighted to hear from anyone wishing to share their data with us with this aim in view.

References

Brummitt, R. K., 1992. Vascular plant families and genera. — Royal Botanic Gardens, Kew. [vi] + 804 pp.

Emmet, A. M., 1992. Life history and habits of the British Lepidoptera. *In*: Emmet, A. M. & Heath, J. (eds.), The Moths and Butterflies of Great Britain and Ireland, 7(2): 61–300. — Harley Books, Colchester.

Karsholt, O. & Razowski, J., 1996. The Lepidoptera of Europe: a distributional checklist. — Apollo Books, Stenstrup. 380 p.

Kartesz, J. T., 1994. A synonymized checklist of the vascular flora of the United States, Canada, and Greenland. — Timber Press, Portland. 1. Checklist. lxi + 622 p.; 2. Thesaurus. vii + 816 p.

McGugan, B. M. (ed.), 1958–1966. Forest Lepidoptera of Canada recorded by the Forest Insect Survey. Canadian Department of Agriculture Publication 1034: 1–6 (1958, [vol. 1]); Canada Department of Forestry Forest Entomology and Pathology Branch, Bulletin 128: 77–281 (1962, vol. 2 — Prentice, R. M., ed.); Canada Department of Forestry, Forest Entomology and Pathology Branch, Publication 1013: 283–543 (1963, vol. 3 — Prentice, R. M., ed.); Canada Department of Forestry Publication 1142: 546–840 (1966, vol. 4 — Prentice, R. M., ed.).

Nye, I. W. B. (ed.), 1975–91. The generic names of moths of the world 1: 568 p. (Noctuoidea (part.) — Nye, I. W. B., 1975); 2: xiv + 228 pp. (Noctuoidea (part.) — Watson, A., Fletcher, D. S. & Nye, I. W. B., 1980); 3: xx + 243 pp. (Geometroidea — Fletcher, D. S., 1979); 4: xiv + 192 pp. (Bombycoidea to Zygaenoidea — Fletcher, D. S. & Nye, I. W. B., 1982); 5: xv + 185 pp. (Pyraloidea — Fletcher, D. S. & Nye, I.W.B. 1984); 6: xxix + 368 pp. (Microlepidoptera — Nye, I. W. B. & Fletcher, D. S., 1991). — British Museum (Natural History)/The Natural History Museum, London.

Robinson, G. S., 1998. Bugs, hollow curves and species-diversity indexes. — *Stats, American Statistical Association* 21: 8–13.

Robinson, G. S., Ackery, P. R., Kitching, I. J. & Beccaloni, G. W., in prep. Hostplants of the moth and butterfly caterpillars of America north of Mexico. Ca. 400 p.

Sattler, K., 1967. Ethmiidae. Microlepidoptera Palaearctica. 2. — Verlag Georg Fromme & Co, Wien. xvi + 185 p. 106 pls.

Scoble, M. J. (ed.), in press. Geometer moths of the world: a catalogue (Lepidoptera, Geometridae). — CSIRO Publications, Melbourne.

Scott, J. A., 1986. The butterflies of North America. — Stanford University Press, Stanford, xii + 583 p.

Silva, A. G. d'Araújo, Gonçalves, C. R., Galvo, D. M., Gonçalves, A. J. L., Gomes J., Silva, M. do Nascimento, Simoni, L. de, 1968. Quarto catálogo do Insetos que vivem nas plantas do Brasil, seus parasitos e predadores 2(1) — Ministério da Agricultura, Rio de Janeiro. — xxvii + 622 p.

Tietz, H. M., 1972. An index to the described life histories, early stages and hosts of the Macrolepidoptera of the continental United States and Canada. — A. C. Allyn, Sarasota, Florida. — vol. 1: 1–536, vol. 2: 537–1041.

Yunus, A. & HO, T. H., 1980. List of economic pests, host plants, parasites and predators in West Malaysia (1920–1978). — [*Bull. Malays. Dept. Agric.* 153.]: i–iii + 1–538. (Ministry of Agriculture, Malaysia).

Zhang, B. C., 1994. Index of economically important Lepidoptera. — CAB International, Wallingford. — 599 p.