Darwin Initiative for the Survival of Species

Final Report

DAISY - Digital Automated Identification of Insects: a way to circumvent the Taxonomic Impediment to monitoring biodiversity, and thus implementing the CBD.

Ian Gauld
Department of Entomology
The Natural History Museum
Cromwell Road
London SW7 5BD
United Kingdom

Darwin Initiative for the Survival of Species

Final Report

CONTENTS

Page

1. Darwin Project Information

Project title:

DAISY -Digital Automated Identification of Insects: a way to circumvent the Taxonomic Impediment to monitoring biodiversity, and thus implementing the CBD.

Country:

Costa Rica

Contractor:

Department of Entomology, The Natural History Museum, London

Project Reference No:

Unknown

Grant Value:

£142,894.00 (agreed award)

Starting/Finishing Dates:

August 1 1997 – January 31 2000; duration 2.5 years

2. Project Background/Rationale

Location and circumstances

The project was conducted jointly between The Natural History Museum, London (NHM) and the Universidad de Costa Rica (UCR). From the latter institution two departments were involved, Biology and Computing Sciences. The lead scientist for the project software development was Dr Mark O'Neill who worked as an independent consultant, for the entire duration of the project, through the NHM. Dr Ian Gauld of the NHM was project leader, Professor Paul Hanson of the Escuela de Biología, Universidad de Costa Rica, was primarily responsible for the Costa Rican co-ordination and arranged for the procurement and preparation of the many insect specimens required. Professor Juan Carlos Briceño, Escuela de Informatica, Universidad de Costa Rica, was the lead Costa Rican computer specialist.

A small number of other collaborators, from both Costa Rica and the United Kingdom, made different inputs at different times during the work. Most notable of these was Señor Ignacio Solis, formerly a computer-science student at UCR who, following graduation, obtained a place at the University of California to study for a higher degree. His programming work on the DAISY front end system undoubtedly helped him obtain this placement and much of the current DAISY front end visuals are attributable to his efforts.

The problem that the project aimed to address

The inventory and monitoring of biological diversity (Article 7) is a critical part of the implementation of the Convention for Biological Diversity. In practice such activities are usually undertaken using a very few groups of organisms (such as birds or amphibians) which are of both high conservation interest and are readily recognisable. There are several disadvantages to using such organisms. First, they are generally high in the food chain, and respond more slowly to environmental perturbation than do invertebrates or micro-organisms. Second, they have a long mean generation time, which also lengthens their response time to such perturbations. Third, many are rare, and sampling can be labour intensive, damaging and difficult.

Many groups of insects do not share these disadvantages. They are often relatively basal in food chains and are the organisms that more charismatic megafauna consume; they often have a very short man generation time and sometimes have two or three (or more) generations in one year; they are generally present in quite large numbers, so destructive sampling is unlikely to have any major impact on the population. Many are monophagous or oligophagous, so thus are potentially good indicators of overall species-richness. Using modern collecting techniques such as Malaise traps, they can be easily and almost continually sampled, and thus enable assessment to be made of how a particular event or treatment impacts on the environment. Using such indicators there are a variety of other activities that may be undertaken. For example, assessment of the “greenness” of coffee fincas in El Salvador is currently being undertaken in part by using insect samples (see Darwin-funded El Salvador project – http://www.nhm.ac.uk/botany/coffee/ projectmain.html).

However, the single, most important factor that prevents the use of insects in biodiversity inventory and monitoring projects is the great difficulty non-specialists have in identifying them. Traditional identification keys can only be used by specialists with access to a large reference collections and specialist libraries. Such facilities are costly to establish and both expensive and difficult to maintain, especially in tropical countries where the climate accelerates decomposition of biological material. This barrier to progress has been referred to as the “Taxonomic impediment”. Various options exist to circumvent this barrier, but one of the most innovative is the application of computer vision techniques to produce an automated identification system. This project was focused on developing such a system, a system that requires very little specialist knowledge by the user.

Evidence of commitment from local partner

The objective of this Darwin-funded project was to develop, in collaboration with the Universidad de Costa Rica, a computerised system for the identification of insects. This was achieved. Initially parasitic wasps were selected as target organisms, but during the course of the project other groups of insects were also used to assess the applicability of the system to a range of different groups. The results have been very encouraging, and an advanced system has been installed at UCR, a duplicate is in the NHM, London, and the software developed has been made available to the National Biodiversity Institute in Costa Rica (INBio).

3. Project Summary

Original purpose and objectives

- to develop an operational automatic identification system that will allow non-specialists to identify species of a target group of insects, and thus permit wide participation in processes designed to inventory and monitor biodiversity.
- to develop the skills necessary in the partner country that will enable our collaborators to use the software to create automatic identification systems for other organisms.
- to demonstrate to the broadest possible audience a novel British developed technique for overcoming the taxonomic impediment to the implementation of Article 7.
- to provide a simple and cost effective method that will allow wide accessibility to diagnostic image-based specimens held in a UK collection.

Modifications during project period

Overall, the project was realised as planned, although we experienced major problems with specimen pose adjustment. Circumventing these problems did delay software development in the middle of the project, but these were successfully overcome well before the project conclusion. Minor modifications were made throughout, as this was a development project with strong research and technology components. For example, a hardware system based on a mounting a digitising camera on separate microscope (as originally proposed) was not adopted because technological advances offered alternatives. Instead we purchased from the Cambridge-based engineering company *Moritex*, a highly sophisticated digital camera system with all necessary microscope functions and illumination system built in. Such a system proved very satisfactory, after a robust, vibration-free mounting system was developed for it in Costa Rica.

Articles of the CBD to which project relates

The project was focused on Article 7, the identification and monitoring of components of biological diversity (notably species). The project also supports Articles 6 (developing national biodiversity strategies, through addressing the obstacle of limited taxonomic capacity at a national level), 8 (gathering of data necessary for effective selection of areas for *in situ* conservation), 10 (recognising components of biological diversity for sustainable use), 12 (research and training in the identification of biological diversity and its components) and 16 (access to and transfer of technology relevant to the conservation of biological diversity).

Results

All the objectives listed above were achieved. A prototype DAISY system was installed and partially developed in the Escuela de Biología at the Universidad de Costa Rica, with a duplicate at The Natural History Museum in London. At UCR the installation occupies a room adjacent to the insect collections, and at the conclusion of the project a fully operational system was in place. Discussions are now underway in preparation to installing the software at the National Biodiversity Institute (INBio) in Costa Rica. As already noted, the results have been very encouraging. Extensive tests clearly indicated the feasibility of the DAISY idea, demonstrated its utility as a practical tool for automating expert identification of insects, and showed how it could be applied to a range of taxonomic groups that, with special reference to the needs of Article 7, would support implementation of the CBD.

During the course of the project, the following specific results were obtained:

- System size scaled up to accommodate 300+ taxa, without adversely effecting the throughput or accuracy of the system. Although this may sound rather trivial, it involved major programming developments. The system now implemented is, at least theoretically, infinitely scalable – that is it can be expanded to accommodate realistically large data sets. The prototype could not, as we found increasing the data set beyond about 50 taxa resulted in both a great increase in computing time and a decrease in identification accuracy.

- Palaearctic ceratopogonid biting midges were identified with 90% success, based on images of wings. It should be noted that all test data were taken from specimens not included in the training data-sets. This level of accuracy compares with success rates of < 20% correct identification of the same material by human non-specialists (test carried out with 2nd year biological science undergraduates at Sheffield University, UK).

- In very large, and taxonomically difficult groups of insects with relatively little information in their wings, such as some bees and ichneumonid wasps, identification accuracy, using a “first past the post system”, was sometimes as low as about 70%. However, using DAISY as a screening engine (i.e. to rank similarities of the unknown to training sets), we were able to reduce the probable identify of unknowns in speciose groups from one in fifty, to one of five parasitic wasps (*Enicospilus* spp.) and to one of two or three bumblebees (*Bombus* spp.). This means that even for very difficult taxonomic groups DAISY is a valuable identification aid, as it narrows down to a few possibilities the identity of an organism. Supplementary characters may then have to be examined for a final identification, but DAISY removes the need to navigate long and difficult traditional dichotomous keys.

- DAISY managed to separate effectively two mosquitoes, *Culex pipiens molestus* and *Culex pipiens pipiens*, based on wing venation alone. Mosquito specialists generally use behavioural (as opposed to taxonomic features) to separate these two subspecies.

- Most recently DAISY has been used to identify moths. In particular Costa Rican hawk moths have been identified with accuracy greater than 80%. The use of the DAISY system for moths and butterflies is obviously a desirable development but has necessitated on-going software changes. The two most obvious needs were for the incorporation of full-colour (DAISY was base on one channel from an RGB input), and for the use of some form of scaling to measure real size (DAISY standardised images for purposes of comparison without recording the magnification necessary to obtain standardisation). Both these improvements have been incorporated into the system since the termination of the Darwin Grant.

In addition, it was also demonstrated that:

- Because DAISY is a generalised image recognition algorithm its potential uses extend far beyond insect species recognition from morphological characteristics. For example it has been shown the system can identify (and track) meso-scale tropical storm systems (e.g. hurricanes) using NOAA and other weather imagery.

- The system can identify heliconiid butterflies and certain fungi from DNA sequence data and/or protein or DNA fingerprints derived from gels.

- The flexibility of the DAISY system is such that it can be successfully integrated with database backends (e.g. D.H. Janzen's Costa Rica caterpillar rearing database, A. Pittaway's Hawk-moths of the Western Palaearctic). This enabled confirmation that the system can be cross-linked with existing state of the art databases to provide a user the means to access a wide range of information about species identified.

4. Scientific Training and Technical Assessment

The system comprises two modified computers, Internet connection and a high quality digitising, magnifying camera with appropriate hardware. Installed on the system is the software, the development of which constituted the single largest part of this project. The software has been extensively tested in Costa Rica, and in the UK, where a somewhat similar system is installed in the Department of Entomology at The Natural History Museum. As each routine was developed, it was extensively tested using images collected from especially prepared specimens. The technical details relating to software development are given in a separate section (below).

Scientific research

Much of the non-software development involved the preparation of a variety of (permanent) test data sets. These comprise sets of images of the wings of various insects. One of the major image sets used for testing were the wings of *Enicospilus* species, a large and difficult genus of nocturnally active wasps belonging to the family Ichneumonidae (as envisaged in the original proposal). The preparation of test sets involved a considerable amount of field work, most conducted at Zurquí de Moravia, a site on the edge of Braulio Carrillo National Park that known to be particularly rich in species. Sampling was conducted from dusk until midnight using specially constructed light traps made to order for the work. These and all other field collecting equipment are deposited in Costa Rica and are now used routinely by students at the Universidad de Costa Rica. Using freshly preserved material collected at light, permanent slide-mounted test specimen sets were prepared, by removing the right fore wing and mounting the wing on a microscope slide in Balsam. The test sets and associated specimens are now deposited at The Natural History Museum, representing scientific vouchers of international significance, but a large number of other specimens (circa 500) were also collected and identified during the course of the project. These specimens are deposited in the Museo de Insectos, Universidad de Costa Rica, and some material has been placed in the collections of the Instituto Nacional de Biodiversidad (INBio).

Although the early focus of the work was wasps, it became apparent that the DAISY system has a wide application, and to investigate this we decided to experiment with other data-sets. Several of these were pre-existing data sets in the NHM or were specially prepared by our UK technician, but a new set was acquired by Dr O'Neill and a photographic technician (M. Denos) in Costa Rica, using fresh, locally collected specimens of hawk-moths (Sphingidae) present in the INBio collection. We opted for this as older museum specimens have often faded, or have their wing scales abraded, which unnecessarily impairs performance of DAISY (although DAISY can cope with such suboptimal specimens, such imperfections are not desirable qualities in the standard reference data-base). The hawk-moth data set has become the standard demonstration set, and features in the TV coverage of the project.

Training and capacity building

As noted above, it was necessary to create test data sets by making preparations of *Enicospilus* wings. Two Costa Ricans were trained in the techniques necessary for making permanent slide mounts.

During the course of this project two visits were made by Costa Ricans to London to participate in software development and system testing. As part of the second visit in February 1999 the three Costa Rican scientists present attended a specially commissioned course at Manchester University. This intensive one-day course, organised by Dr Tony Lacey and Dr Neil Thacker, from the school of Image Science and Bio-Medical engineering explained the latest developments in TINA, a state of the art set of programming libraries used for machine vision and image processing in experimental biophysics. Knowledge of these programming routines was of considerable importance, because it was initially thought that some of the routines could be used in the DAISY programme. Although they were eventually not used, familiarisation with the TINA system prevented duplication of effort and led to a fruitful exchange of ideas.

Technical computational details

At the start of the DAISY project there were no generic scaleable systems appropriate for a task such as insect recognition, although concurrently other systems have been developed. However, none of these is a truly automated identification system designed for a non-expert user. For example, the system built to recognise bee species from their wing venation implemented by Wittmann and colleagues at the University of Bonn in Germany requires considerable user intervention. Although this system is capable of high accuracy (e.g. correct classification rates in excess of 95% for certain bee groups, for example *Colletes* species) it requires the user to accurately locate a large number of vein intersections in the wings. This means that the system is not easily usable by the non-expert unfamiliar with the particular venational terminology (and several alternative systems of venational terminology exist for insects), and it is relatively labour intensive and slow to do single determinations. Furthermore, the Wittmann system, at least in its earlier forms, is simply a sequence of a number of pre-existing computer codes derived from the work of the Photogrammetric and Remote Sensing communities. Consequently, the system does not lend itself to generic scaleable recognition in the way that the DAISY system does.

DAISY represents a completely novel approach to insect identification. Basically a system was envisaged that accepts an image and identifies it, without user intervention, by comparing it with other images stored in a reference database. In its initial form (e.g. Weeks, P.J.D., Gauld, I.D. Gaston, K.J. & O'Neill, M.A. 1997. Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research*, 87: 203-211.), DAISY was based closely on the PCA-based facial recognition systems developed by Matthew Turk and Sandy Pentland at Massachusetts Institute of Technology in the early 1990s. The prototype DAISY system was essentially a re-implementation of the Turk and Pentland principal component analysis (PCA) face classifier. However, a major modification to the approach advocated by Turk and Pentland was the building of individual classifiers for each morph-class (e.g. in the case of DAISY, insect species). There were two major reasons for this, first to ensure that DAISY was scaleable by having the "identification engine" access an expandable set of concatenated reference images. Second, by having concatenated images, effectively constructs embracing the observed variation in the morph-class, we were able to recognise variable objects such as insect species. Previously, such systems worked only for invariant objects, such as certain unalterable reference points derived from a human face or fingerprints.

It became clear that if computer-aided taxonomic (CAT) systems like DAISY were to be a practical proposition the amount of computation associated with adding (a) new species to the system or (b) increasing the number of training examples for a given species must be limited. In classical PCA the entire set of training examples is (non-linearly) transformed to a single reduced dimensionality space. While this is not a problem with small numbers of species and closed training sets, it becomes untenable for recognition systems that may contain thousands of individual species and tens of training examples per species because:

- Every time a new species is added to the system, or a new example is added to a species training set, the reduced dimensionality form of the entire training set (*N* species * *M* examples) needs to be recomputed. As this is *O*(*M***N*^2) the size of the problem rapidly become intractable.

- The matrix algebra associated with computing the principal components is not easy to parallelise given a loosely coupled MIMD array (e.g. a local area network of workstations running NT or LINUX).

Although the prototype DAISY approach worked relatively well (correct classification in >95% of all identifications for 5 closely related species of Costa Rican parasitic wasps, and >90% in the case of some 30 species of Palaearctic ceratopogonid biting midges), following initial testing as part of the Darwin Initiative funded project two further problems with the PCA approach were recognised. First, the PCA approach is very computationally intensive, which means even the modified PCA approach adopted by the prototype DAISY project actually scaled relatively poorly. Second, PCA implicitly assumes that the distribution functions in morph space of objects that are to be recognised are essentially linear. This is clearly not the case for biological objects such as insect wings.

In order to overcome these deficiencies, during the first year of the Darwin-funded DAISY project a second version of DAISY was implemented which used the Lucas *n*-tuple nearest neighbour classifier (NNC) as opposed to PCA in order to compute the affinity of unknowns to training sets. This classifier has several advantages over the earlier PCA approach. First, NNC is very simple. The unknown is assumed to be in the same class as the training example to which it has the highest affinity. Second, NNC is capable of dealing with non-linear distributions of training objects in morph-space. Third, it scales linearly with increasing training set size and number of species. Fourth, it may easily be implemented in hardware (should the need arise). Fifth, it is amenable to MIMD parallelisation of the classification process over a network of interconnected workstations. Finally, it is capable of supporting sophisticated training set optimisation algorithms with a minimum of extra computational overhead.

Tests conducted using the ceratopogonid midge and parasitic wasp data sets analysed previously by the prototype showed that the NNC version of DAISY performed at least as well as the earlier PCA versions (on the criterion of correct identifications), and with a throughput speed which was at least an order of magnitude better on the same hardware.

The final DAISY system consists of classification engines (FLORETS) which are based on the Lucas *n*-tuple NNC classifier together with a number of other components developed or adapted during the course of this project, which facilitate data input and the dissemination of information about the objects classified. These components are:

1. DAISY FRONT END (DFE) – a X11R6 based GUI front end built using the FSF Gnome and GDK+ libraries. DFE is a tool which can be used to capture imagery, mark the boundaries of objects (such as insect wings) within the imagery which are to be

identified, and perform a selection of image-processing operations on the input imagery (e.g. contrast enhancement, centering, reflection, etc.).

2. IPM: ipm is a front end to the floret classification engine that transforms input objects into a standard pose (e.g. invariant to rotation and scale).

3. A virtual HTML generator, which turns a list of probable identifications generated by floret into a link page of HTML. It then launches a standard Web Browser (e.g. NETSCAPE) in order to display this information.

In practice, the exemplar DAISY system described above has been able to discriminate similar species rich in visual information, such as hawkmoths (Sphingidae). Species of the genus *Xylophanes* have consistently been identified to species with a very high degree of accuracy (approaching 100%). In its current form, DAISY uses the modified n-tuple NNC to build an ordered list of distances from the unknown in morph space. This means that the system is able to fail gracefully. If it cannot make an identification to species, it is almost always able to say that the unknown *X* is one of (say) ten species. This means that the system has a useful screening function in the case of sibling species complexes, and species swarms.

For example, extensive tests conducted during this project with 55 species of Costa Rican parasitic wasps in the genus *Enicospilus* (which contains complexes of extremely similar and morphologically rather variable species possessing very similar wing venation) have shown that DAISY is almost always able to say the unknown *X* is one of four or five possible species. This is extraordinary as non-specialists have extreme difficulty identifying these organisms. The screening function of DAISY will save a lot of time when dealing with speciose tropical biota, as it greatly reduces the number of species that have to be considered. It reduces the identification burden on the expert taxonomist because screening can be performed by relatively inexperienced personnel, such as parataxonomists.

5. Impacts

Accomplishment of project purpose and goals

Evidence for the accomplishment of the project's purpose and goals is documented elsewhere in this report (section 3).

Contribution to CBD articles

Article 7 As stated in section 3, the project was focused on Article 7, the identification and monitoring of components of biological diversity, with special reference to species. In addition, it also made a contribution to the following articles, as also explained in Section 3:

Article 6 (developing national biodiversity strategies, through addressing the obstacle of limited taxonomic capacity at a national level);

Article 8 (gathering of data necessary for effective selection of areas for *in situ* conservation);

Article 10 (recognising components of biological diversity for sustainable use);

Article 12 (research and training in the identification of biological diversity and its components); and

Article 16 (access to and transfer of technology relevant to the conservation of biological diversity).

Training and capacity building

The impacts on training and capacity building are described in sections 3 and 4, and summarised in section 6 below.

Impacts on collaboration

Impacts on collaboration between the UK and Costa Rican partners are described under sections 2, 3 and 8.

Beneficiaries of the project

The main beneficiaries of the project, in the short term, are UCR and INBio, and consequent development of better capacity for the Costa Rican government to discharge its duties under the CBD (with special reference to Article 7). Longer term, such is the generality of the research and technical developments undertaken during the project, potentially all signatories to the CBD as well as and local people could benefit indirectly. This assumes that further progress is made in developing DAISY as an automated identification tool of wide applicability and use for biodiversity evaluation and monitoring.

Publicity

The DAISY system has received local attention in Costa Rica where an article appearing in the leading local newspaper, *La Nacion*. Most recently, the BBC programme *Tomorrow's World* (May 30, 2001) presented a feature on the DAISY system, demonstrating how it was helping to identify insects in Costa Rican National Parks.

6. Outputs

Publications:

Gauld, I.D., O'Neill, M.A. & Gaston K.J. 2000. Driving Miss DAISY: the performance of an automated insect identification system, pp 303-312. *In:* Austin A.D. & Dowton, M (eds.) *Hymenoptera: Evolution, Biodiversity and Biological Control.* 468 pp. CSIRO. Canberra.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. & Gauld, I.D. 1999. Species-identification of wasps using principal component associative memories. *Image and Vision Computing*, 17: 861-866.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. & Gauld, I.D. 1999. Automating insect identification: exploring the limitations of a prototype system. *Journal of Applied Entomology*, 123: 1-8.

DAISY web site – http://chasseur.usc.edu/pups/projects/daisy.html

Darwin Initiative Standard Output Measures:

Output measure 2 (training).

Number of Masters qualifications attained: Initially we had envisaged a Costa Rican MSc student, but the most suitable collaborator (and the extraordinary programming requirements of the project necessitated the highest calibre students) was a final year honours student. He spent two periods of approximately 6 weeks in the UK.

Output measure 4a (training).
Number of undergraduate students receiving training: 3 undergraduates worked at various times on aspects of this project. The most outstanding was Señor Ignacio Solis, who upon completion of work on DAISY as his final year's project obtained placement for a higher degree at the University of California in the USA

Output measure 6b (training).
Number of training weeks not leading to formal qualification: A considerable period, in total some 16 weeks, was spent by Professor Paul Hanson and myself training Costa Ricans in the preparation of wings slides of the various species, that were used for image capture.

Output measure 8 (research).
Number of weeks spent by UK project staff on project work in host country: During the project the PI spent almost 9 months in Costa Rica, both on this and other collaborative projects. Whilst it is difficult to precisely define how much time was spent I estimate over 8 weeks were spent on the DAISY project, both in training collaborators and in collecting specimens. Stemming from this project, over 500 hundred specimens have been identified to species, and the majority have been deposited in the Costa Rican National Collection at INBio, with representative identified material in the collection of the Museo de Insectos, Universidad de Costa Rica. The Principal computer scientist in the project, Dr O'Neill spent a bout 3 weeks in Costa Rica, installing and developing software on the UCR system.

Output measure 11a (research).
Number of papers published in peer-reviewed journals and books: Several papers resulting from this work have been referred to above. These record the technical development of the computing and the results obtained using various regimes. As part of the collecting undertaken for the project several new species have been discovered. These have been included with large amounts of species from other sources and the results have been published as two taxonomic monographs:

Gauld ID, 2000. Ichneumonidae of Costa Rica, 3. *Memoirs of the American Entomological Institute*, 63: 1-453.

Gauld ID, (ed.) 2002. Ichneumonidae of Costa Rica, 4. *Memoirs of the American Entomological Institute*, 66: 1-768.

Output measure 11b (research).
Number of papers published elsewhere: In addition to the traditional publications listed above a DAISY website has also been established.

Output measure 12a (research).
Number of computer-based databases established containing species information and handed over to host country: 2. All databases of captured images are available in Costa Rica.

Output measure 13b (research).
Number of species reference collections enhanced: Three principal collections have been enhanced during this project, INBio an the Museo de Insectos in Costa Rica, and the collection

of the NHM in London. Representatives of new species, including most usually the primary types have been deposited in the Costa Rican national collection at INBio.

Output measure 15a (dissemination).
Number of national press releases in host country: 1. An articles appeared in *La Nacion*, a major daily newspaper in Costa Rica.

Output measure 18b (dissemination).
Number of national TV programmes in UK: 1, the Tomorrow's World programme as detailed above.

Output measure 20 (physical).
Estimated value of physical assets handed to host country: The computer equipment and software will remain at the UCR. This has an estimated value of £12,000 at time of purchase, although of course the DAISY software would greatly exceed this if marketed.

7. Expenditure

Table 1:

The total awarded by the Darwin Initiative was £142,894.00

Summary and variations

8. Operation and Partnerships

Initially this was a partnership between the NHM London and two schools of the Universidad de Costa Rica, Biology and Informatics/Computing Sciences. As the project progressed, it became clear that it was necessary to expand the project to include a more high profile group of organisms. Hawkmoths were chosen, and so collaboration was sought and obtained from INBio. Their collections were the source of most material of this group, as most of the specimens have been collected very recently and are in near perfect condition. The fading and abrasion that occurs in older museum specimens reduces the value of specimens for the purposes of training set generation.

There is no other comparable project running in Costa Rica, but we did liaise closely with the Informatics division of INBio. No international organisations participated in the project, and no such co-operation or partnership was ever anticipated.

Funding is being sought by UCR to build on the DAISY system installed in the Escuela de Biologia, to develop an application to identify insects of quarantine importance, tephritid fruit flies. Dr O'Neill is corresponding with Dr Erik Mata, Head of Informatics at INBio, about the installation of a working DAISY system in INBio.

9. Monitoring and Evaluation - Lesson Learning

The system was tested iteratively throughout the project, leading to a series of improvements to the basic algorithms. Many of these improvements are outlined in two published papers, Weeks *et al*. (1999) and Gauld *et al*. (2000). A comparative study was undertaken using 2nd year undergraduates versus a naïve DAISY user. The results showed that a very much higher level of correct identification could be achieved by DAISY than by undergraduates using traditional keys.

To detail milestones in algorithm development, particularly one such as DAISY, where months were spent by the computer scientists writing code which would then be tested extensively, is difficult. Certain problems were identified in the course of the work, the most notable being pose adjustment – that is ensuring the test image for recognition is perfectly aligned with the reference images. Over the course of software development vast numbers of tests were made and all major problems have been overcome. The fact that the system could be used by undergraduates with no previous experience of computer sciences (such as Anna Watson, see below), demonstrates the progress made towards having an easily usable system.

The best external evaluation of the DAISY system was made during an Honours project by Anna Watson, University of Bangor, during 2001–02. She found that it worked very well and, as indicated under Section 11, it is now anticipated that she will be involved in further developments, successful research grant applications permitting.

10. Darwin Identity

Darwin logo is used on the opening screen of the DAISY program, and this is apparent wherever and whenever the program is used. Although DAISY is a distinct entity, this acknowledgement will be carried in future versions. The Darwin Initiative has also been acknowledged in the three published papers derived from the project, and on the project website. More generally, within Costa Rica there is now wide awareness of the Darwin Initiative (e.g., through the article published in *La Nacion*), and at the Universidad de Costa Rica and at INBio. Awareness and identity of the Darwin Initiative in the UK was boosted by the BBC TV *Tomorrow's World* programme, screened on 30th May 2001.

11. Leverage

Based on the success of the project and the ability to demonstrate the feasibility of automated insect identification, various initiatives and partnership are currently being sought that will fund further development. In the UK an undergraduate, Ms Anna Watson, has used the Darwin-funded DAISY system to explore the possibility of identifying live moths at a light trap, working in North Wales. The results were very encouraging, her third year project report was highly rated, and she recently graduated with first class honours. The NHM is now actively seeking a BBSRC studentship in bioinformatics to pursue this idea, by engaging Ms Watson to work on tropical hawkmoths and on-line identification as PhD study. She is currently in Central America capturing images of moths, in the hope of taking the concept forward. Success at identifying live insects at a light trap via the Internet would be a major step in the automation of direct diversity monitoring – an entirely novel approach with the potential to have a profound influence on the future of tropical biology, and biomonitoring in general.

Discussions are also in progress between the NHM and Cornell University regarding the application of DAISY to the identification of published illustrations of insects. Such a programme would bring an entirely new dimension to releasing information from the vast but largely inaccessible heritage that exists in the scattered and fragmented literature on biological diversity. In parallel, a currently separate programme is already underway to digitise and database the *Biologia Centrali Americana*, a lavishly illustrated 70 volume work first published over 100 years ago. The *BCA* still forms the baseline for much of our knowledge of Central American biodiversity. This digitisation project is headed by the American Museum in New York and the Smithsonian Institution in Washington, with the NHM as a partner. The possibilities for synergy between these two programmes are very exciting.

12. Sustainability and Legacy

The current project has shown that an automated insect identification system, an utterly novel concept, is both practical and viable. Thanks to Darwin Initiative funding, the DAISY ideal has been developed into an adaptable working system, and it has also demonstrated its potential for a range of uses. Development has continued and partners have kept in touch. Interest has been expressed by INBio to use DAISY as a tool to allow identification of a range of Costa Rican insects and provide instant access to information held in their own databases. Scientists at the Universidad de Costa Rica are interested in developing databases to use DAISY to screen for agricultural pests. The use of DAISY was also favourably discussed at a recent international meeting on the implementation of the Global Taxonomy Initiative (GTI) held in Costa Rica for Central American states.

DAISY can thus be seen as a significant conceptual contribution towards alleviating the “taxonomic impediment” that hinders our understanding of the planet’s biodiversity, and it has every possibility of being developed as one of the most powerful tools for overcoming this obstacle. Its great strength lies in its capacity to capture expert knowledge, not only of living experts, but also of deceased taxonomists (or classifiers of stars, cultural artefacts, pot shards, handwriting or what have you) where they have left a legacy of well-classified and documented collections of specimens, images or other artefacts. To achieve true sustainability the system will have to be developed further. The impetus given by this Darwin Initiative grant to the continuing development of the DAISY concept has been fundamental to developing the ability to reach that ultimate goal. The legacy of this work can be immense.

One major difficult that must be faced for developing this type of system is the need to secure enough funding to populate large scale databases with sets of images based on digitising expertly identified specimens (be they wasp wings, moths, stars or storm-clouds). This is the key step to capturing sufficient specialist knowledge to create a truly ‘expert system’. This type of activity is not viewed as scientifically interesting by research councils, or even by many other bodies concerned with R&D. In effect we need a separate funding body that appreciates the need to capture expert knowledge in this way.

13. Value for Money

Given that the project was completed using less than the original budget, its specific successes, and the potential for further development and application of the DAISY system, the project represents excellent value for money.

Ian D. Gauld
Department of Entomology
The Natural History Museum
Cromwell Road
London SW7 5BD
United Kingdom

Revised report: 30 July 2002