The BioNET-INTERNATIONAL GROUP FOR COMPUTER-AIDED TAXONOMY (BIGCAT)

PROCEEDINGS OF INAUGURAL MEETING

held at

The University of Wales

Cardiff

2-3 July 1997

Published by the BioNET-INTERNATIONAL Technical Secretariat (TECSEC)
Bakeham Lane, Egham, Surrey , TW20 9TY

PROCEEDINGS
OF THE
INAUGURAL MEETING OF THE
BioNET-INTERNATIONAL
GROUP FOR
COMPUTER-AIDED TAXONOMY
(BIGCAT)

BioNET-INTERNATIONAL GROUP FOR COMPUTER-AIDED TAXONOMY

BIGCAT

Compiled by
David Chesmore and Lynne Yorke
School of Engineering
University of Hull
and
Paul Bridge of CABI Bioscience
UK Centre Egham
Edited by Simon Gallagher (TECSEC)

Published by
The Technical Secretariat (TECSEC)
of BioNET-INTERNATIONAL

CONTENTS

SESSION 1 AUTOMATED SYSTEMS

SESSION 2 KEY SYSTEMS

SESSION 3 VR AND TRAINING

WORKING GROUP REPORTS

FOREWORD

The BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT) was established at an inaugural workshop convened jointly by the Technical Secretariat of BioNET-INTERNATIONAL and the Wales Centre for Biodiversity, and kindly hosted by Professor Mike Claridge at the University of Cardiff on 2nd-3rd July 1997. It brought together some 30 researchers in the fields of biology, ecology, mycology, nematology, entomology, computing, film making and engineering. Its purpose was to review recent developments in computer-aided taxonomy (CAT) and assess prospects of further application of existing CAT and of developing new systems to meet BioNET-INTERNATIONAL's needs. This volume includes the main papers presented and the discussions which followed.

The concept of BIGCAT emerged following a presentation of BioNET-INTERNATIONAL at a seminar of the Welsh Pest Management Forum held on 25th-27th April 1997 at the University of Wales Field Station at Gregynog, where the need for user-friendly, computer-aided and automated taxonomic systems by the developing country LOOPs of BioNET-INTERNATIONAL was emphasised. It was the enthusiastic response of the delegates at the Seminar that led to the subsequent workshop.

BIGCAT is now firmly established and is developing its own agenda with four main fields of interest, *viz.* automated systems, key systems, education and training systems and one that deals with BioNET-INTERNATIONAL's special needs. Present membership is, by chance, solely from the UK, but the Group is anxious to extend its membership especially in Europe to include all interested specialists in fields relevant to its remit.

An excellent opportunity to attract new members to BIGCAT will occur at the Second BioNET-INTERNATIONAL Global Workshop to be held at the University and the National Museum & Gallery of Wales at Cardiff on 22nd-29th August 1999 where a number of CAT systems will be exhibited by experts at the Workshop's Technology Fair.

Tecwyn Jones

ACKNOWLEDGEMENTS

The Technical Secretariat of BioNET-INTERNATIONAL (TECSEC) wishes to record its gratitude to all who attended the Inaugural Workshop at Cardiff on 2nd-3rd July 1997 and contributed so effectively to the discussions. The School of Pure & Applied Biology (PABIO) which provided facilities, equipment and hospitality is especially acknowledged as is the support and assistance given by the host, Professor Mike Claridge and Dr Hassan Dawah. The Chairpersons, Dr Dave Chesmore, Dr Bill Hominick, Dr Marion Edwards and Dr Bill Parker are thanked for their leadership and guidance.

The publication of these Proceedings was funded with the financial assistance of the United Kingdom Department for International Development (DFID).

INTRODUCTION

By Professor Tecwyn Jones
Director Technical Secretariat and Chairman
BioNET-INTERNATIONAL Coordinating Committee

When BioNET-INTERNATIONAL was conceived and promulgated in 1991, as a global network of sub-regional LOOPs in developing countries and a few technical support LOOPs in the developed world, one feature that attracted favourable attention was its finite time-scale for achieving objectives. With its four LOOP work programmes, i.e. for information and communications, training of taxonomists and technicians, rehabilitation and development of collections and records, and the development and use of new technologies, it was predicted that the objective of realistic self-reliance in taxonomy in the developing country sub-regions was achievable within ten years of the beginning of operations.

Analysis of the tasks to be accomplished, and the actual and potential resources available to do so, suggested that any significantly shorter perspective was unrealistic, whilst any significantly longer projection was not only unnecessary but might also cause loss of impetus and focus and other drawbacks symptomatic of very long-term programmes.

By what may seem a happy coincidence the ten year time-scale of BioNET-INTERNATIONAL proved to be very acceptable to the developing country governments involved in establishing, operating and sustaining LOOPs. It was also seen as appropriate by the national scientists charged by their governments with developing the level of taxonomic capacity needed for sub-regional self-reliance. Equally importantly perhaps, the ten-year projection provides an acceptable funding scenario for bilateral and multi-lateral development agencies. It provides them with the opportunity for effective short-term (3-5 years) inputs in the form of projects to support the ongoing longer-term programme within the clearly focused and permanent infrastructure of the sub-regional LOOPs.

Whatever its appeal, however, the finite time-scale of BioNET-INTERNATIONAL was the minimum realistic perspective that emerged from a broad analysis of the task to be accomplished by the LOOPs and the actual and potential resources available to do so. Whilst these included financial resources, such as the likely commitment of national governments to sustaining the LOOPs and their South-South programmes and the prospects of donor funding for North-South cooperation, it was the assessment of scientific resources that was the most critical at the time.

It was evident then, as it is now, that if the world’s taxonomic capacity remains at its present level and if, as a consequence of this, the advancement of taxonomic knowledge continues at its present slow pace, this fundamental resource will never catch up nor keep pace with the world's demand for it. At the present rate, even if, through BioNET-INTERNATIONAL, parity is achieved between the taxonomic capacity of the developed and developing worlds, this will still be inadequate, especially in developing countries, to provide the taxonomic support needed by national programmes for sustainable agricultural development, conservation and sustainable use of the environment and biodiversity.

Against this background it was recognised from the outset that, in building taxonomic capacity, LOOPs should not only be concerned with training taxonomists and taxonomic technicians but also with making taxonomy available to, and useable by, a very much wider community, and that to do this a variety of new technologies needed to be acquired, developed and used. The question arising was: Are such technologies available and/or capable of being developed and how and where could they be applied?

The answer was extremely encouraging. The advent of computerised identification keys and semi and/or fully automated systems for the identification of organisms, clearly made BioNET-INTERNATIONAL's objectives realisable within an acceptable time-scale and, if further developed and used, could indeed make taxonomy available to and useable by a community of hitherto unimaginable dimensions. User-friendly computer-aided technologies such as interactive identification keys, digital imaging, audio and photo sensing systems, neural networks, etc. and digital identification systems make it possible for applied scientists not least plant protection and plant quarantine personnel, ecologists, environmentalists, agronomists, agricultural and forestry specialists and biotechnologists, etc. to become adequately competent in the taxonomy of the organisms of particular concern to them. Furthermore, the most sophisticated electronic keys used by specialists can be modified and simplified to suit the needs of those further down the taxonomic stream, whilst pictorial keys offer appropriate identification tools in many circumstances and, where taxonomic niceties are involved, molecular systems are available to solve problems.

It is self-evident that taxonomy must become as automated as possible and as quickly as possible if humankind is to become capable of inventorying the vast array of organisms with which it shares this planet, and the BioNET-INTERNATIONAL community needs automated systems to achieve its objectives. It is committed to using what is available at every opportunity and to developing and adapting technologies to suit its particular needs. It was for this very purpose that the BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT) was formed in 1997.

BIGCAT's purpose is to keep abreast of all new technologies, use, modify and adapt existing technologies as necessary to suit the network's needs and develop new technologies, with the ultimate objective of automation. Given necessary funding for development and field and laboratory testing, it has within its corporate competence the potential to make substantial contributions to the advancement of taxonomic knowledge and progress in inventorying the world's biodiversity, monitoring it and understanding the relationships of the species within it.

This first BIGCAT publication gives an insight into the interests and capabilities of the group, and as membership extends, as is the intention, its expertise, capacity and potential will be substantially augmented. BIGCAT is a vital programme within BioNET-INTERNATIONAL's remit as a source of innovation and application. Its ultimate goal is the down-stream application of the taxonomy produced by the world specialists and of relevant technologies from other sciences that help this application and/or provide new means of distinguishing and understanding the components of the world's biodiversity.

SESSION 1

AUTOMATED SYSTEMS

METHODOLOGIES FOR AUTOMATING THE IDENTIFICATION OF SPECIES

By E.D. Chesmore
Department of Electronic Engineering, University of Hull, Hull HU6 7RX

1. Introduction

The application of computers in systematics can be divided into three broad categories - identification keys, automated identification and implementation of methodologies for taxonomy such as cladistics. All three categories are grouped together under the banner of *computer-aided taxonomy* (CAT). Whilst key systems are well developed for some groups and computers are used widely for cladistics, the concept of automated species identification is still in its infancy with relatively little research effort at present. Recent advances in computing power and signal processing techniques are opening up new opportunities for the development of detection and identification systems for a wide range of taxa. This paper is a review of techniques and methods available for automatically identifying taxa and gives examples from recent research projects at Hull University.

2. Automating Species Identification

The process of species identification by electronic means can be considered to be an application of general pattern recognition in which an unknown is placed into one of a number of possible classes depending on features extracted from measurements on the unknown. Pattern recognition has many applications ranging from handwriting recognition to speech analysis and identification of faults in machinery (condition monitoring). Automated species identification is closely related to many of these applications. Two main levels of automation can be identified - partially and fully automated as described below:

a. **Fully Automated**. Complete identification without user interaction; this requires highly reliable identification with a very low probability of misclassification.
b. **Semi-automated**. This category is perhaps more realistic than a) as it allows prior sorting into higher taxonomic categories such as genera and presents the user with data for further manual identification if required. It is essentially a relaxation of fully automated identification and is therefore more likely to be feasible in the short term.

It is anticipated that semi-automated systems will be the most viable as they allow the user to perform the final identification which is considered to be more acceptable in the short term, as there is a tendency for humans to mistrust computers or consider them as a threat which may result in a potential impediment to CAT. It is therefore expected that semi-automated systems will play an important role in validation of techniques and in obtaining general acceptance of automation.

There is still relatively little research into species recognition; Table 1 gives a summary of recent projects indicating that the potential applications are widespread.

Table 1. Examples of Current Identification Systems.

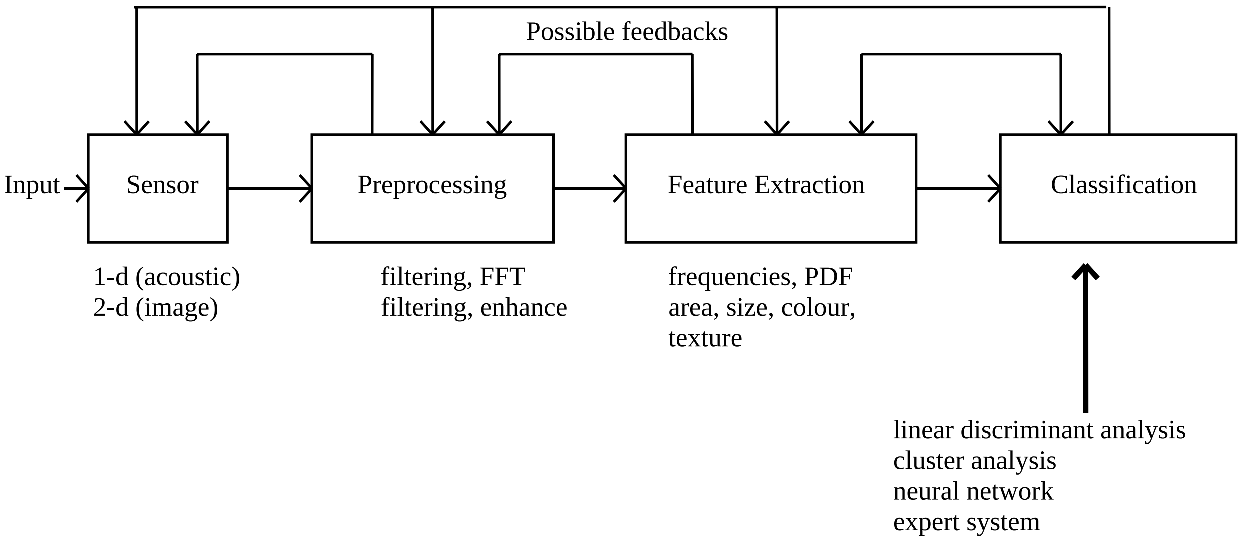
Current Examples	Technique(s)
Fish species (e.g. Mackerel)	Active acoustics (sonar), PDF, cluster, ANN
Orthoptera	Passive acoustics ("listening"), TDSC + ANN
Frogs	Passive acoustics, frequency
Birds	Passive acoustics, frequency
Mosquito species	Passive acoustics (amplitude, frequency)
Flying insects (size and mass estimation only)	Radar, IR Doppler (wing beat frequency)
Lepidoptera	Monochrome & colour image analysis
Phytoplankton	Flow cytometry, image analysis, ANN
Hymenoptera	Monochrome image analysis
Leaf-miners	Monochrome & colour image analysis
Fungal spores	Monochrome image analysis
Plants, weed species	Monochrome & colour image analysis
Nanofossils	SEM + image analysis
Pollen	SEM + image analysis

Notes: SEM = scanning electron microscope
ANN = artificial neural network
PDF = probability density function
IR = infrared
TDSC = time domain signal coding

2.1 Structure of an Identification System.

Figure 1 shows the structure of a generalised species identification system which takes the form of a classical pattern recognition system comprising four functional blocks - sensor, preprocessor, feature extractor and classifier (Shalkoff, 1992). The term *classifier* is used here in a classical engineering sense - a device for assigning features to labels, and should not be confused with the taxonomic meaning of classification. It is evident from Figure 1 that the classification process is sequential in nature but can have feedbacks to allow modification of, for example, features depending on classification results (e.g. to select different features to improve classification). The following sections discuss each functional block in detail.

Figure 1. Generalised Automated ID System.



2.2 Sensor

In order to perform an identification, the object(s) must be sensed in one or more ways. Sensors can be grouped into several categories depending on the group to be identified; this can be seen from Table 1. The categories are:

a. **Images**. Monochrome, colour, infra-red (IR) and thermal IR image can be obtained using cameras, either digital or via photographs scanned into digital form. Scanning electron microscope (SEM) images can also be used for some applications.

b. **Acoustics**. This is divided into active and passive. Active acoustics (sonar) uses reflected acoustic signals from animals such as fish to perform identification (Scalabrin *et al*., 1996; Simmonds *et al*., 1996). Passive acoustics is basically listening and is applicable for many animal groups such as insects (Chesmore, 1997; Chesmore *et al*., 1997; Hagstrum *et al.*, 1990; Shuman *et al.*, 1993), amphibia (Taylor *et al.*, 1996), birds (Mills, 1995) and mammals (e.g. bats and cetaceans). Typical sensors would be high quality microphones or hydrophones.

c. **Radar**. Active radar sends pulses of electromagnetic radiation and analyses returned echoes. Such systems can be used to map insect swarms or in some applications derive body mass, wingbeat frequency and direction for individual insects (Smith *et al.*, 1993). An alternative is harmonic radar where target insects have a microwave diode and small antenna attached to their bodies. The antenna re-radiates energy at a harmonic of the illumination frequency. Harmonic radar has been used to successfully track bees in flight and ground beetle movement.

d. **Infra-red (IR).** IR can take 2 forms - firstly, wingbeat frequency can be obtained using a simple IR illumination and detection system. Secondly, thermal IR images can be taken. Sensors range from simple IR photodiodes to thermal IR imagers (very expensive).

e. **Flow Cytometry**. This obtains scattering properties of single cells such as phytoplankton by using a laser beam and optical sensors to detect side scatter. See Boddy *et al.*, (1994) for more details.

Selection of an appropriate sensor depends mainly on the taxa involved and the application. In some cases, the solution is relatively obvious, for example, acoustic identification of bats, whereas in others there may be a choice. One possible example of the latter is in phytoplankton analysis which could use image processing of individual cells or flow cytometry. Each method has its relative merits (including speed, reliability and cost) which must be considered at the outset. In addition to the selection of sensors, there are a number of considerations to be taken:

a. Images: Pixel size in relation to object size
bits/pixel = no. of levels (B/W or RGB)
memory size (X × Y × bits/pixel × no. of planes)
lighting, colour balance, reproducibility
distortion

b. Signals: frequency response of sensor
sampling criterion (time and/or spatial)
quantisation effects (bits/sample)
signal to noise ratio (SNR) (background noise)

2.3 Preprocessing

Preprocessing usually involves increasing the signal to noise ratio (SNR) of the measurement. Examples include:

- Analogue or digital filters (1-d and 2-d) for noise reduction or interference removal.
- Analogue-to-digital conversion.
- Contrast enhancement, closure of small holes, skeletonisation of images.
- FFT (fast Fourier transform) to convert between time and frequency domains.
- Time domain signal coding (TDSC) symbols.
- Crosscorrelation, autocorrelation to detect periodicity.

2.4 Feature Extraction

The extraction of features is the most important part of the system, inappropriate selection of features will lead to poor classification. Typical features for various sensors are:

a)	Images:	object shape, texture, size, orientation, colour, colour moments, skewness, mean/sd of RGB (red, green, blue) components, edge density, HSV (hue, saturation, value) components;
b)	Acoustics:	dominant frequencies, vocal tract information, temporal structure, amplitude PDFs, TDSC codes;
c)	Radar:	scattering parameters, Doppler shift, polarisation.

2.5 Classification

The final stage of the process is to classify the features, i.e. assign data to one or more prespecified classes based on various features or signal attributes. Classifiers include:

- linear discriminant classifiers;
- cluster analysis (statistical);
- syntactic pattern recognition (formal language grammars);
- artificial neural networks (various forms: RBF, backpropagation, TDNN, self-organising feature maps);
- expert systems (rule-based systems, blackboard systems).

Knowledge-based systems such as expert systems and artificial neural networks have the ability to handle disparate, fuzzy and incomplete data which makes them suitable for automated identification. Knowledge-based systems should be considered as fundamental to the success of any computer-assisted taxonomy system and will form the central core, data being supplied by the user.

3. Example - Recognition of Orthoptera using Bioacoustics

The concepts described in Section 2 will be illustrated by an example of insect species identification using time domain signal coding (TDSC) of bioacoustic signals and artificial neural network classification. Acoustic detection of insects has been carried out for some agricultural applications, particularly pests in grain stores (Shuman *et al.*, 1993; 1997), however, species are not identified. The work described here aims to go one stage further, i.e. to identify species using

British Orthoptera as a readily available test set. The technique and results are described in detail in: Chesmore *et al.* (1997). The sensor, preprocessor, feature extractor and classifier for this application are given in Table 2. The sensor, being a microphone (or output from a recording) will not be discussed further except that it's frequency response must be matched to the insects' frequencies. The other 3 sub-systems are described in more detail in the following sections.

Table 2. Pattern Recognition Sub-systems for Example System.

Sensor	Audio microphone (or output from pre-recorded sounds)
Preprocessor	16-bit analogue to digital conversion (Soundblaster card) followed by time domain signal coding (TDSC) – generates stream of codewords
Feature Extractor	Co-occurrence matrix (A-matrix) of pairs of codewords
Classifier	Perceptron neural network (others include MLP, self-organising maps, blackboard systems)

3.1 Preprocessor

Before any signal analysis can take place, the acoustic signals must first be converted into digital form. In this application the sounds are sampled at 44kHz, 16-bits per sample using a program called Goldwave V3.03 on a PC via the line input of a Soundblaster card. Subsequent signal analysis is carried out using Matlab.

The signal analysis preprocessor used in this application is a technique known as time domain signal coding (TDSC), also known as time-encoded signals which was developed in the 1970's by King and Gosling (1978). It has subsequently been used in a number of applications including acoustic condition monitoring of machinery (Lucking *et al.*, 1994) and heart sound analysis and defect identification. TDSC characterises any bandlimited signal by its "shape" between successive real zeros (termed an epoch); generally, this shape is taken between actual zero-crossings. Each epoch is described in terms of its duration in samples (D) and shape (S) usually taken as the number of minima or signal energy as indicated in Figure 2 which shows a 10 sample epoch with 2 minima (D=10, S=2). The number of possible D-S combinations (symbols) is termed the natural alphabet which can often be non-linearly mapped onto a smaller symbol set to give signal compression. In the original speech application, the coded symbols were transmitted and used to regenerate the speech signal at the receiver thus providing digital speech transmission at substantially reduced data rates.

3.2 Feature Extractor

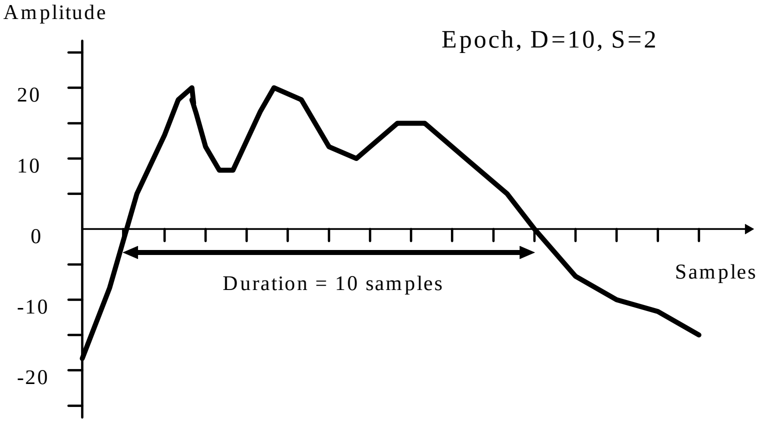
TDSC can be described as the concatenation of a signal's D-S symbols, i.e. it produces a sequential stream of symbols and one analysis method is to examine the occurrence of pairs of symbols over time to give a histogram, A, which describes the number or proportion of symbols i and j occurring in succession, i.e. the number of times i is followed by j by a lag L. A 2-dimensional histogram, the A-matrix, can be formed, expressed mathematically as:

$$a_{ij} = \frac{1}{(N-L)} \sum_{n=L+1}^{m=N} x_{ij}(n)$$ where L = lag; $x_{ij}(n)$ = 1 if t(n) = i and t(n-L) = j (0 otherwise)

and t(n) = n^{th} TES symbol

This fixed size histogram with time-invariant dimensions is the feature set used for classification purposes. The entry at position (i,j) represents the number (percentage) of occurrences of the TDSC symbol pair i and j where j is delayed relative to the first (in epochs). In this application, a lag of 1 epoch is used; multiple lags may also be employed giving rise to multi-dimensional matrices. The A-matrix is independent of any gain factors if the input signal has no dc component and is therefore insensitive to relative energies of different segments of the signal.

Figure 2. Definition of a TDSC Epoch.



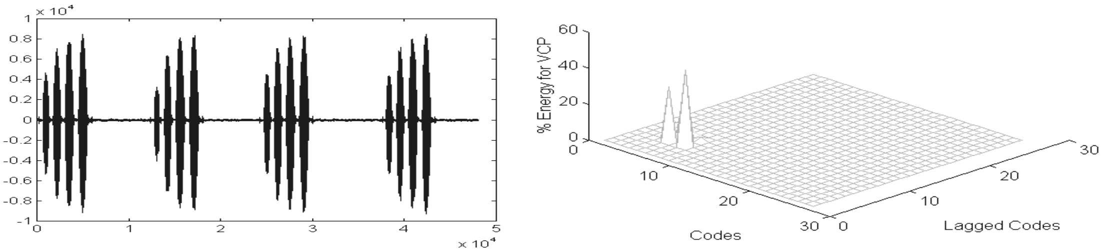
3.3 Classifier

The A-matrix is unique for each sound (call, syllable, etc) and forms the basic feature for pattern classification using an artificial neural network (ANN). ANNs are now widely used in many classification and identification problems as they can be trained, are good at handling fuzzy and disparate data and are able to perform non-linear discrimination. There are many forms of ANN which can be divided into supervised (requires training) and unsupervised classification (no training). The majority of classification methods currently used are based on multilayer perceptrons (MLP) using backpropagation for training. More recently, self-organising feature maps have been investigated with some success.

3.4 Results for British Orthoptera

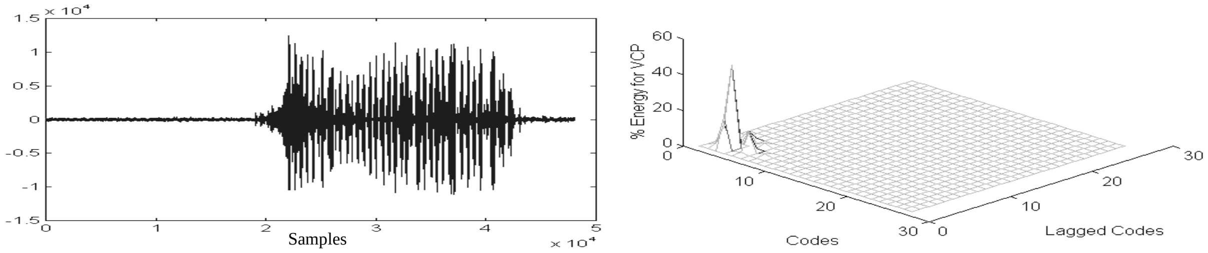
Table 3 lists the 25 species used in 1 set of tests. The sounds were derived from a widely available audio cassette (Burton and Ragge, 1987) available as an accompaniment *to The Grasshoppers and Allied Insects of Great Britain and Ireland* (Marshall and Haes, 1988) and digitised as previously described. TDSC symbols for up to 2 seconds of sound for each species. The subsequent A-matrices were used to train a simple single layer Perceptron consisting of 784 inputs (1 for each location in the A-matrix) and 25 output neurons, 1 for each species. Figures 3 and 4 show time plots and A-matrices for 2 species - *Gryllus campestris* and *Chorthippus albomarginatus*. Since the perceptron only performs linear discrimination (n-dimensional hyperplane), these results suggest that TDSC is a good pre-processor providing wide separation of sounds which would have similar spectra. This is evident when the example A-matrices in Figures 3 and 4 are examined.

Figure 3. Time Domain Waveform and A-Matrix for *Gryllus campestris.*



Time Domain Waveform **A-Matrix**

Figure 4. Time Domain Waveform and A-Matrix for *Chorthippus albomarginatus.*



Time Domain Waveform **A-Matrix**

Once trained, the system was tested with new sounds; Table 3 shows identification results with various levels of added noise to simulate response to poorer conditions. Each entry is an average of 1000 normally distributed random A-matrices (zero mean, unity variance) added to the A-matrices which simulates Gaussian white noise over the whole frequency spectrum. It is evident from Table 3 that identification is very high (99-100%) under low noise conditions with the exception of OR06 (*Metrioptera brachyptera*) and that misidentification is zero until 30% noise is added. The latter is a fundamental requirement for an automated system. However, it is important to note that some species exhibit a very rapid decline in identification accuracy (OR01 - *Meconema thalassinum* and OR13 - *Gryllotalpa gryllotalpa*). Both species have characteristically low dominant frequencies and this may contribute to confusion. However, it is known that the former species produces substrate-based sounds and would not be detected in a bioacoustic survey. Variations in classifier using MLPs and expert system identification have been assessed, all with reasonable results (Chesmore, 1997; Chesmore *et al.*, 1997).

4. Applications

Potential applications include:

Insect Counting. Little research has been carried out in this potentially important area. Gonzales (1986) developed a pilot image processing system for identifying insects from agricultural surveys in an effort to speed up the often time consuming sorting process. It has been suggested that systems of this nature could aid considerably in sorting from large catches even if the sorting

Table 3. Orthoptera Species Identification Accuracy with Added Noise.

Latin Name	English Name	ID Code	Noise Level (%) 1	3	5	10	20	30	40	50
Meconema thalassinum	Oak Bush-cricket	**OR01**	99.9	83.5	54.2	24.3	8.7	5.3	4.7	4.3
Tettigonia viridissima	Great Green Bush-cricket	**OR02**	100	100	100	100	96.7	89.3	82.3	76.8
Decticus verrucivorus	Wart-biter	**OR03**	100	99.6	94.8	76.8	63.8	60.6	58.6	53.6
Pholidoptera griseoaptera	Dark Bush-cricket	**OR04**	100	100	100	99.9	94.1	86.8	79.3	72.0
Platycleis albopunctata	Grey Bush-cricket	**OR05**	100	100	99.4	90.1	71.1	66.9	59.0	57.9
Metrioptera brachyptera	Bog Bush-cricket	**OR06**	85.6	65.5	58.8	54.3	54.3	52.0	52.0	47.9
Metrioptera roeselii	Roesel's Bush-cricket	**OR07**	100	100	99.9	94.2	65.0	45.1	29.1	21.6
Conocephalus discolor	Long-winged Cone-head	**OR08**	100	100	100	100	99.1	95.1	87.4	79.0
Conocephalus dorsalis	Short-winged Cone-head	**OR09**	100	100	100	100	100	99.7	98.4	95.5
Acheta domesticus	House-cricket	**OR10**	100	100	100	100	99.4	94.8	84.5	76.0
Gryllus campestris	Field-cricket	**OR11**	100	100	100	100	100	100	100	100
Nemobius sylvestris	Wood-cricket	**OR12**	100	100	100	100	99.1	93.8	84.1	73.8
Gryllotalpa gryllotalpa	Mole-cricket	**OR13**	100	84.9	63.7	34.2	16.5	9.3	7.6	6.0
Stethophyma grossum	Large Marsh Grasshopper	**OR14**	100	100	100	100	93.4	85.7	76.1	73.6
Stenobothrus lineatus	Stripe-winged Grasshopper	**OR15**	100	100	99.7	89.3	74.9	61.5	58.9	52.2
Stenobothrus stigmaticus	Lesser Mottled Grasshopper	**OR16**	100	99.9	98.0	84.9	68.0	63.8	59.2	59.4
Omocestus rufipes	Woodland Grasshopper	**OR17**	100	86.4	74.4	63.4	58.2	52.8	50.2	45.8
Omocestus viridulus	Common Green Grasshopper	**OR18**	100	98.1	86.7	59.7	47.0	38.1	35.4	32.8
Chorthippus brunneus	Field Grasshopper	**OR19**	100	100	99.9	95.6	80.7	71.5	62.8	62.9
Chorthippus vagans	Heath Grasshopper	**OR20**	100	100	99.9	91.3	77.1	67.7	63.7	60.5
Chorthippus parallelus	Meadow Grasshopper	**OR21**	100	100	100	100	99.8	97.5	92.7	81.8
Chorthippus albomarginatus	Lesser Marsh Grasshopper	**OR22**	99.8	86.1	74.1	62.3	54.5	56.4	53.5	54.5
Euchorthippus pulvinatus	Jersey Grasshopper	**OR23**	100	100	100	99.7	94.1	84.6	78.5	71.9
Gomphocerippus rufus	Rufous Grasshopper	**OR24**	100	99.6	92.0	65.8	48.9	39.6	36.7	30.6
Myrmeleotettix maculatus	Mottled Grasshopper	**OR25**	100	100	100	100	100	99.8	98.9	96.6
	Mean identification (%)		99.4	96.1	91.8	83.4	74.6	68.7	63.7	59.5
	Mean false identification (%)		0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.3

process only identifies to order or genus. Such pre-sorting could reduce the identification time by an order of magnitude. This also links to biodiversity assessment and pest monitoring (see below).

Biodiversity Assessment. Riede (1993) suggested that since many rainforest species produce sounds, it may be possible to use acoustic analysis for monitoring fauna. Riede used Orthoptera for more rapid biodiversity estimation in a tropical lowland forest in Ecuador, and Oba (1994, 1995) used bird song as a measure of the "natural sound diversity" in Japan. Until very recently, it has only been possible to identify species manually from recordings which is both costly and time consuming; the development of automated identification systems will speed up the process and lead to continuous real-time monitoring.

Ecological Monitoring. Potential applications for ecological monitoring are diverse and include recording the occurrence of call types and correlating with environmental conditions, long term continuous monitoring and determination of bird species for species-specific bird strike avoidance (bird scarers) in airports. It is also theoretically possible to identify and monitor individuals in populations of some taxa (e.g. birds).

Pest Monitoring. As noted in Section 3, many animal pests, particularly insects, can be detected by their sound production. In the USA, Shuman *et al.* (1993) used acoustics for detecting beetle larvae in rice grains. Hagstrum *et al.* (1990) used similar techniques for monitoring of *Rhizopertha dominica* (Coleoptera: Bostrichidae) in wheat kernels. It is theoretically possible to detect and, more importantly, identify many different insect and animal pests in a variety of agricultural and horticultural environments although very little work has been done to date.

5. Conclusions

This paper has briefly covered methodologies for automated species identification. To date, little research has been carried out in this potentially important research area. Much work needs to be done not only in developing techniques and methodologies but also in increasing the awareness of researchers of the potential benefits of automated identification.

The paper illustrated some of the concepts using an example of ongoing research at the University of Hull in bioacoustic signal recognition. This work currently concentrates on insects but will shortly be expanded to include birds and, hopefully, bats.

6. References

Boddy, L., Morris, C.W., Wilkins, M.F., Tarran, G.A. and Burkill, P.H. (1994) Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry* **15**, 283-293.

Burton, J.F. and Ragge, D.R. (1987) *Sound Guide to the Grasshoppers and Allied Insects of Great Britain and Ireland.* Harley Books, Great Britain.

Chesmore, E.D. (1997) Neural networks and expert systems for automated insect identification, BES Ecological Computing Group Workshop on *"Machine learning methods for ecological applications",* Manchester Metropolitan University.

Chesmore, E.D., Swarbrick, M.D. and Femminella, O.P. (1997) Automated analysis of insect sounds using TESPAR and expert systems - a new method for species identification. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 273-287.

Gonzales, N.N. (1986) Insect Identification using Template Matching: a Pilot Study. Master's Thesis, New Mexico State University.

Hagstrum, D.W., Vick, K.W. and Webb, J.C. (1990) Acoustic monitoring *of Rhizopertha dominica* (Coleoptera: Bostrichidae) populations in stored wheat. *Journal of Economic Entomology* **83**, 625-628.

King, R.A. and Gosling, W. (1978) Time-encoded speech. *Electronics Letters* **15**, 1456-1457.

Lucking, W.G., Darnell, M. & Chesmore, E.D. (1994) Acoustical condition monitoring of a mechanical gearbox using artificial neural networks. *IEEE Conference on Neural Networks*, Florida, USA.

Marshall, J.A. and Haes, E.C.M. (1988) *Grasshoppers and Allied Insects of Great Britain and Ireland*. Harley Books, Great Britain.

Mills, H. (1995) Automatic detection and classification of nocturnal migrant bird calls. *Journal of the Acoustic Society of America* **97**, 3370-3371.

Oba, T. (1994) Sampling methods for the study of the natural sound environment in Japan: consideration of the sample time unit. *Natural History Research* **3**, 27-32.

Oba, T. (1995) What is the natural sound diversity? A consideration for the local natural amenity. *Natural History Research* **3**, 173-185.

Riede, K. (1993) Monitoring biodiversity: analysis of Amazonian rainforest sounds. *Ambio* **22**, 546-548.

Scalabrin, C., Diner, N., Weill, A., Hillion, A. and Mouchot, M. (1996) Narrowband acoustic identification of monospecific fish shoals. *ICES Journal of Marine Science* **53**, 181-188.

Schalkoff, R. (1992) *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley & Sons, Inc., New York.

Shuman, D., Coffelt, J.A., Vick, K.W. and Mankin, R.W. (1993) Quantitative acoustical detection of larvae feeding inside kernels of grain. *Journal of Economic Entomology* **86**, 933-938.

Shuman, D., Weaver, D.K. and Mankin, R.W. (1997) Quantifying larval infestation with an acoustical sensor array and cluster analysis of cross-correlation outputs. *Journal of Applied Acoustics* **50**, 279-296.

Simmonds, E.J., Armstrong, F. and Copland, P.J. (1996) Species identification using wideband backscatter with neural network and discriminant analysis. *ICES Journal of Marine Science* **53**, 189-195.

Smith, A.D., Riley, J.R. and Gregory, R.D. (1993) A method for routine monitoring of the aerial migration of insects by using a vertical-looking radar. *Philosophical Transactions of the Royal Society of London, B* **340**, 393-404.

Taylor, A., Grigg, G., Watson, G. and McCallum, H. (1996) Monitoring frog communities: an application of machine learning. *Eight Innovative Applications of Artificial Intelligence Conference*, AAAI Press.

DAISY: AN AUTOMATED INVERTEBRATE IDENTIFICATION SYSTEM USING HOLISTIC VISION TECHNIQUES

By M.A. O'Neill
Digital Vision, Edmonds Court, Didcot, Oxfordshire OX11 8QY

I.D. Gauld
The Natural History Museum, Cromwell Road, London SW7 5BD

K.J. Gaston
Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN

P.J.D. Weeks
21 Outram Road, Oxford OX4 3PD

1. Introduction

Taxonomy has largely failed to deliver what much of society requires from it – the means to identify the species comprising life on Earth. Despite more than two centuries of taxonomic activity, the overwhelming majority of terrestrial species of organisms have never been formally described and documented, and most of the described species are so poorly characterised they are unrecognisable to all but a few specialists with access to historical reference collections. Few taxonomic experts exist to recognise these taxa, and dichotomous printed keys are often impossible to use without both access to a reference collection and a knowledge of specialist terminology. Even if the literature to identify an organism exists biologists may be unable to use it (Gauld, 1986). Printed keys have been augmented by the use of computerised multi-access keys, beginning with text based keys (Pankhurst, 1978) and culminating recently in multimedia works such as CABIKEY (White and Scott, 1994), but these still rely on the ability of the user to discriminate subtle pictorial information. Using computers to present taxonomic characters, while relying on users to compare specimens, images, or illustrations represents a failure to utilise the potential offered by information technology. Automated identification methods based on image analysis have occasionally been used (see Weeks and Gaston, 1997), but no large scale taxonomic identification system potentially scaleable to hundreds of species has ever been achieved. The objective of the DAISY (Digital Automated Identification SYstem) project is to use computer vision technology to mechanise the discrimination process and isolate it from observer errors.

DAISY works with insect wing images because they are two dimensional, and less likely to present a computer vision system with pose-related visual recognition problems than three-dimensional objects. The approach has been motivated by the recent progress which has been made in human face detection and recognition using fuzzy template matching techniques based on decomposing a training set (a series of examples of the object to be recognised) into a set of orthogonal eigenimages (principal components) (Turk and Pentland, 1991; Pentland *et al.*, 1994). Unknown objects can then be classified by seeing how well they correlate with an optimal linear combination of the principal component eigenfunctions. This sort of technique operates directly on the grey levels of an image and is distinct from the traditional feature-based approach in which items of interest within images are represented in terms of semantic data such as the distances and angles between intensity anomaly features. While the grey level representations may be sensitive to variation in illumination and pose, they have a clear advantage in that information may automatically be derived from the statistical structure of the imagery and therefore the difficult problem of selection of individual features is eliminated (Valentin *et al.*, 1994). Such procedures may readily be automated – and scaled – while feature based methods must rely on interactive feature selection and measurement.

2. Materials

Our research has been focused in two areas, five species of two genera (*Pimpla* and *Neotheronia*) of parasitic wasps (Hymenoptera: Ichneumonidae), and 50 species of two genera (*Culicoides* and *Forcipomyia*) of small biting midges (Diptera: Ceratopogonidae). These were selected because both families comprise large numbers of species which often differ from each other in subtle ways, and consequently are notoriously difficult to identify.

3. Data Acquisition

Details of specimen preparation, image acquisition and pre-processing operations are given by Weeks *et al*. (1996), but briefly, images of slide mounted wings were captured using a Kontron ProgRes 3000 colour CCD camera, mounted on a Zeiss Stemi SV11 Apo stereomicroscope and stored to disk in a personal computer running an image analysis software package (KS400, Kontron Elektronik GmbH). Red and blue components of the imagery were discarded, and the green component was transposed to a grey scale image. This image was then corrected for shading, using a shading reference image. Light intensity and other optical parameters were kept constant during the imaging process. After capture, the basal part of the wing was cropped out. The images which were captured at 768 x 400 pixels were reduced in size so the cropped area is 100 x 100 pixels. This was undertaken to make principal component analysis of the imagery more tractable. Prior to undertaking principal component analysis, the images were processed in order to enhance interspecific differences and intraspecific similarities in both venation and pigmentation patterns. These operations included the removal of background and local feature enhancement by application of “top hat” filtering to enhance wing venation patterns followed by a Gossip 3 x 3 image blur.

4. The Image Recognition Algorithm

The classifier within the DAISY system operates as an associative memory. Images which are similar to those used to train a particular classifier (conspecifics) have high correlation coefficients when compared with that classifier, whilst less similar images have lower correlation coefficients. The classifier is trained by exposing it to a set of images (the training set images) which are selected to span the expected (intraspecific) variation of the object to be recognised. Principal component analysis is then used to generate a series of *n* eigenimages from the original *n* training set images. These images express the variation inherent in the original training set images, but unlike the input imagery they have the property that:

$$\langle \mathbf{A}_i | \mathbf{A}_j \rangle = 0 \quad (i <> j) \qquad (1)$$
$$\langle \mathbf{A}_i | \mathbf{A}_i \rangle = \mathbf{k}$$

Where: $\mathbf{k}$ is a constant ($\mathbf{k}$ = 1 if images are orthonormal),
i and j are image indices,
$\mathbf{A}_i$ and $\mathbf{A}_j$ are eigenimages,
$\langle \mathbf{A}_i | \mathbf{A}_j \rangle$ is the pixel dot product of images i and j.

The orthogonal nature of the eigenimages means that we can decompose any unknown image X in terms of the eigenset $\{\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_i, \dots \mathbf{A}_n\}$. Thus:

$$\mathbf{C}_j = |<\mathbf{X}|\mathbf{A}_j>| \qquad (2)$$

Where: $\mathbf{C}_j$ is the spectral strength of the j^{th} member, of the eigenset $\{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_i, \dots \mathbf{A}_n\}$,
$<\mathbf{X}|\mathbf{A}_j>$ denotes the pixel dot product of **X** and $\mathbf{A}_j$.

This yields a spectral strength vector $\{\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_i, \dots \mathbf{C}_n\}$, which expresses **X** in terms of a linear sum of the eigenimages of the eigenset **A**. This spectral strength vector may then be used to reconstruct the unknown image **X** in terms of the eigenset $\{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_i, \dots \mathbf{A}_n\}$:

$$\mathbf{X}' = \textstyle\sum_i \mathbf{C}_i|\mathbf{A}_j> \qquad (3)$$

Where: $\mathbf{X}'$ is the reconstructed image,
$\mathbf{C}_i$ is the i^{th} spectral strength coefficient,
$\mathbf{A}_i$ is the i^{th} eigenimage.

The affinity of **X** for the PCA associative memory $\{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_i, \dots \mathbf{A}_n\}$ may then be estimated by computing the correlation coefficient A_f between $\mathbf{X}'$ and **X**:

$$A_f = corr(\mathbf{X},\mathbf{X}') \qquad (4)$$

Where: A_f is the affinity of X for $\{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_i, \dots \mathbf{A}_n\}$
(0 ~ no affinity; 1 ~ images identical),
X is the unknown image,
$\mathbf{X}'$ is a reconstruction of unknown image linear sum of eigenimages $\{\mathbf{A}_1,\mathbf{A}_2,\dots,\mathbf{A}_i,\dots\mathbf{A}_n\}$.

There are a variety of possible choices of correlation function (*corr*) which could potentially be used to correlate the unknown image and its projection in principal component space. A set of these (metric distance, Malahobis distance, and a modified cross-correlator) were tested. Turk and Pentland (1991) used a simple metric distance function when computing the distance between an unknown image and "face space". However, this metric gave poor results with visually very similar insect wings. We found that metrics based on either non-parametric statistics (Kendall-t) or a modified cross-correlation measure were more effective.

The form of PCA used in the DAISY classifier also differs from that described by Turk and Pentland in another important respect. In order to identify given faces these authors effectively identify the spectral strength vector associated with a given face F, where the corresponding training set contains all the faces the system currently "knows". This has the disadvantage that the training set, which is potentially very large, has to be recomputed each time a new face is added to the system. In contrast the DAISY system is modular. The DAISY system has a separate classifier for each object it needs to identify. This means that the classifier for one object can be modified without having to recompute the eigensets of the other classifiers within the system, and that further classifiers for new objects (i.e. species) can be added to the system without directly altering any existing classifiers.

The use of modular classifiers within the DAISY system means that adaptive (self-learning) classifiers which are capable of modifying their own training data may be implemented relatively easily. Although the cost in terms of computational resources of making an identification is slightly higher than that for the Turk and Pentland implementation, the modularity, scalability and possibilities for adaptive learning more than compensate for this.[1]

[1] As Pentland *et al.* (1994) have discussed, linear PCA is restricted to small regions within the object space continuum which can be modelled by linear functions. However, we can easily extend this simple PCA prescription to deal with non-linear (highly 3-dimensional) objects if we do not mind incurring additional computational cost.

5. Species Identification Using DAISY

The viability of the approach outlined was addressed using a database of wing images. Species classifiers were generated for each of five species of ichneumonids (Weeks *et al.*, 1997) and for each of fifty species of ceratopogonids (Weeks *et al.*, submitted). Each classifier was trained on 5-15 wing images of its target species (the training set). Other images were used to generate a set of test images which were presented to the classifiers to be identified. Thus all identifications made by DAISY in our tests were undertaken on images that had not been used to generate the classifier. In tests, the highest correlation was deemed to be the "correct" identification. A major problem for an identification system is recognising that a species submitted to it (an "alien") does not belong to any of the included taxa. This is as much a problem for DAISY as for any other system, for all images are likely to correlate at some level with the classifiers. Thus the performance of classifiers when presented with aliens was assessed by presenting them with images of wings of other species.

The proportion of specimens of each species identified correctly using the DAISY PCA implementation is uniformly high. Using Kendall-t as the metric, in the majority of instances the highest correlations observed were between test images of species *X* and the classifier which has been trained to identify that particular species. In the case of the ichneumonid data set 94 percent of the test images were correctly identified (Weeks *et al.*, 1997), and of the six percent of specimens which were misidentified half were correctly identified when they were re-imaged and re-analysed. The situation with the much larger ceratopogonid data set was broadly similar with 86 percent of the material correctly identified (Weeks *et al*., submitted).

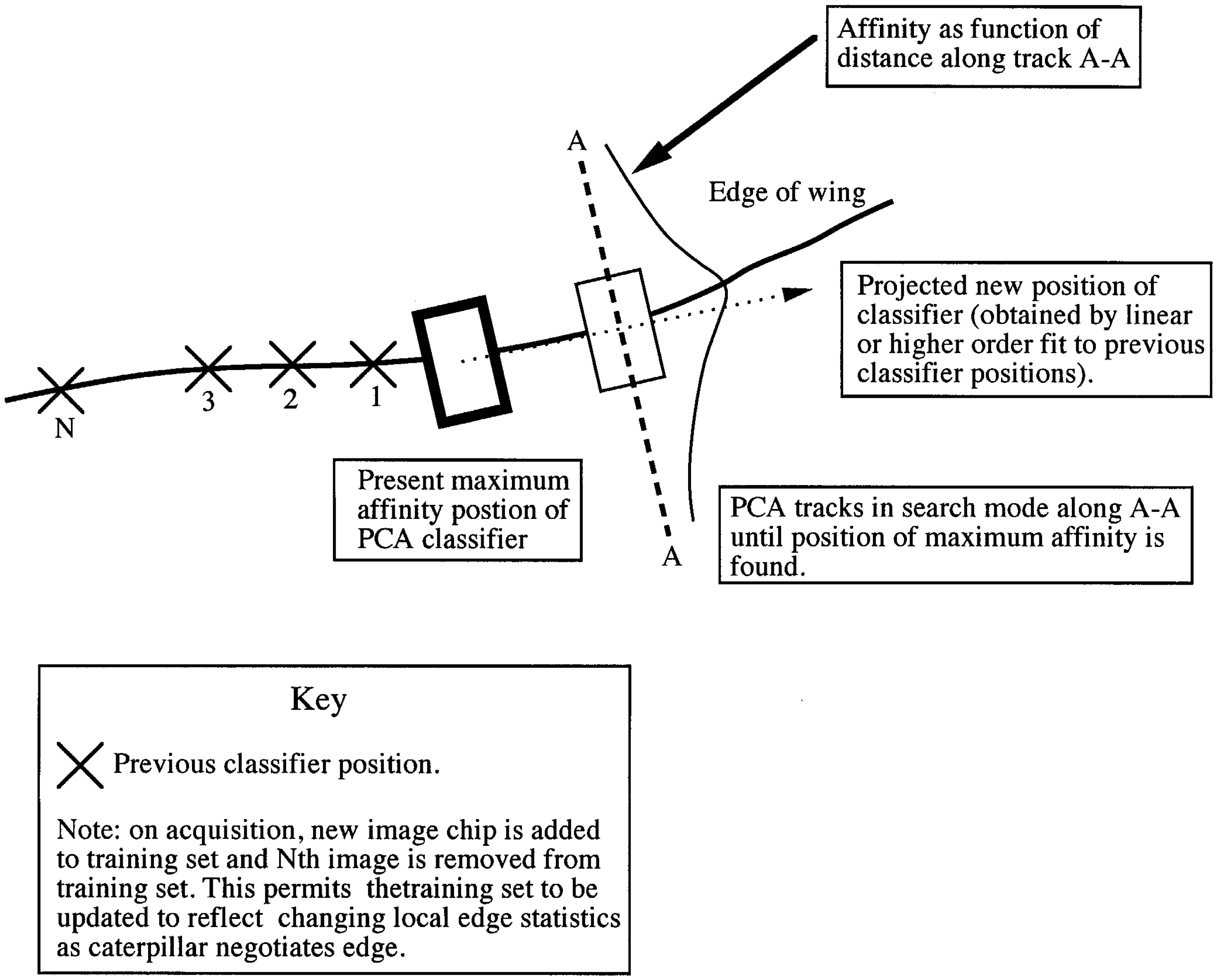
In the case of the ichneumonid data set aliens were all found to have very low Kendall-t correlations with all the classifiers, even in instances where the alien was congeneric with the classifier species. For example, the European *Pimpla hypochondriaca* had a Kendall-t of 0.434 with the Costa Rican *Pimpla croceiventris*, considerably lower than values obtained for test specimens of *P. croceiventris* (which were around 0.8). This suggests the possibility of establishing a threshold correlation value, below which no identification will be made.

6. Discussion

The results of the feasibility study are promising as the success rate is high and compares favourably with the results obtained for human face recognition (96%; Pentland *et al.*, 1994). It is likely that non-taxonomists presented with the same material as DAISY would be markedly less accurate in identifying it.

Despite the encouraging results obtained, the present system is far from providing routine taxonomic identification. For its promise to be realised, significant obstacles must be overcome. First, levels of successful identification between 85 and 95 percent may not be adequate, especially if the organism to be identified is a major commercial pest. The fact that about half of the incorrectly identified specimens were subsequently correctly identified when image capture was repeated, suggests that refining this process may provide one means of increasing the number of correct identifications. In this regard, the use of snakes (Ivins and Porrill, 1993; Schabel and Arridge, 1995) (Figure 1) may significantly improve matters. These algorithms automate the operation of separating imaged wings from background noise by identifying the wing envelope. Subsequently, with the use of appropriate image moments, the wing may be transformed to a standard orientation and scale. An alternative method of wing alignment may be afforded using the pairwise geometric histogram (PGH) transform (Ashbrook *et al*., 1995). In this case, a caterpillar algorithm is used to extract the wing venation and bounding envelope. These are then transformed into a pairwise geometric histogram, the form of which is a function of the wing geometry. The histograms for the training set and unknown images are then processed by the DAISY classifiers in the usual manner. The wing envelope data produced by snakes and/or caterpillars may be shape and scale transformed in its own right and thus used as an additional identification cue[2]. Greater control and/or manipulation of data-capture parameters may also be important. For example, the light intensity at which images are captured can influence identifications. Experiments suggest that in conditions of decreasing light intensity, specimens of *Neotheronia mellosa* are increasingly likely to be misidentified as *N. lineata*, and with increasing light intensity as specimens of *Pimpla sumichrasti*. Furthermore, the present implementation of DAISY does not make full use of the RGB (red/green/blue) images captured. Discarding the red and blue fields of the image in the case of some species may well cause misidentifications which would not otherwise occur.

Figure 1. A schematic of the Caterpillar algorithm.



Affinity as function of distance along track A-A

Projected new position of classifier (obtained by linear or higher order fit to previous classifier positions).

Present maximum affinity postion of PCA classifier

PCA tracks in search mode along A-A until position of maximum affinity is found.

Previous classifier position.

Note: on acquisition, new image chip is added to training set and Nth image is removed from training set. This permits thetraining set to be updated to reflect changing local edge statistics as caterpillar negotiates edge.

[2] It should be noted that snakes and caterpillars could also be used to automatically extract wing data to be used with classical methods of wing image analysis such as those advocated by Yu *et al.* (1992).

Second, it is not a trivial problem to distinguish specimens which are visually very similar. At present a test specimen is identified as the species to whose classifier it has the highest correlation, even when it has almost similarly high correlations with other species. This “first past the post” method of identifying specimens is almost certainly sub-optimal. Even if the correct classifier is not the winner (in a first past the post sense), it is possible that the holistic behaviour of the entire classification vector {$\mathbf{C}_1$, $\mathbf{C}_2$, ..., $\mathbf{C}_1$, ..., $\mathbf{C}_n$} may be used to identify the specimen. This would require a meta-classification stage in which a self-organising neural net or a second stage PCA classifier is used to try and identify patterns in the classification vector which are characteristic of a given species. Alternatively, a combination of theorem proven and associative list processor (e.g. the CANTOR system in R-transform mode (O’Neill and Hilgetag, 1998)) could be used to extract (species-specific) sets of (correlation coefficient) ordering rules from the classification vector[3] .

Another way of reducing misidentifications is to use more characters. This may be useful as, although DAISY may misidentify a specimen, the species to which it belongs may, by correlation coefficient, be ranked close behind the “first past the post”. For example, analysis of the ceratopogonid results showed that if one accepts as a correct identification one of the highest five correlations the success rate rises to 94 percent. DAISY has narrowed down the set of species one needs to compare an unknown with from 50 to five. Other character sets could then be requested for final identification. The modular nature of DAISY means that the building of aggregate (multi-character) species classifiers is a relatively easy task.

Third, in order to distinguish alien specimens of species for which no classifier exists, it is necessary to set a threshold for the correlation coefficient, below which a specimen is deemed to be unknown. To date, this threshold value has been set in an arbitrary fashion but more testing is necessary to try to establish the range of values likely to be generated by intraspecific variation. If the threshold value is set too high for *X*, specimens of *X* may be excluded as unknowns, whilst if it is set too low alien specimens may not be screened out as unknowns. It is possible that PCA, neural network, or list processor based meta-classification may provide a route for reducing the magnitude of this problem.

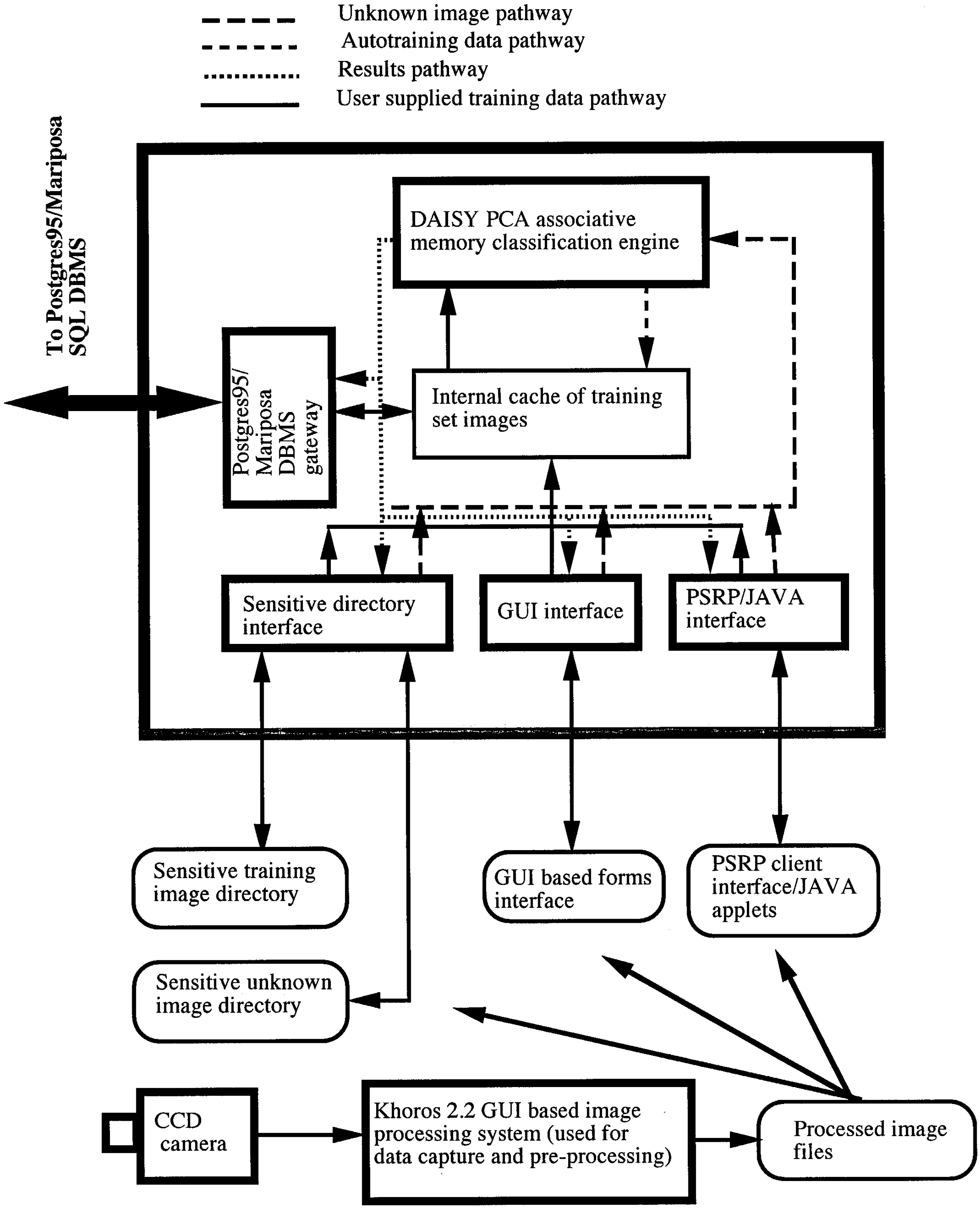
Fourth, it is unclear how best to organise screening of images against large numbers of classifiers. If the classifiers are organised as a single (huge) classification vector, the computational load on the system will increase as a linear function of the number of species for which classifiers exist – and in the upper limit of thousands of species, this approach is likely to be untenable. A hierarchical tree, the branches of which reflect the taxonomy of the organisms for which classifiers exist, is an attractive proposition. A genus classifier could, for example, be trained on a selection of images which are representative of species within that genus. It may then be possible to identify specimens to genera using genus classifiers and then subsequently to species, using the species classifiers of the winning genus classifier. Depending on its success, this approach could be extended to many taxonomic levels, and with the use of robust clustering techniques, such as NMDS (Young *et al.*, 1995) it may be possible to generate genus classifiers adaptively. Alternatively, it may be more feasible to have a series of entirely “artificial” hierarchies that more closely reflect phenetic similarity than phylogenetic affinity, but the predictivity of such constructs would of course be limited as one could not necessarily obtain a generic identification if a specific one were not possible.

[3] These rules (or a subset thereof) may serve as a fingerprint for a given species.

7. Future Developments

In order to enhance the performance of DAISY it is likely that substantial modifications will be made to the basic algorithm described above. These modifications will include: the use of snakes

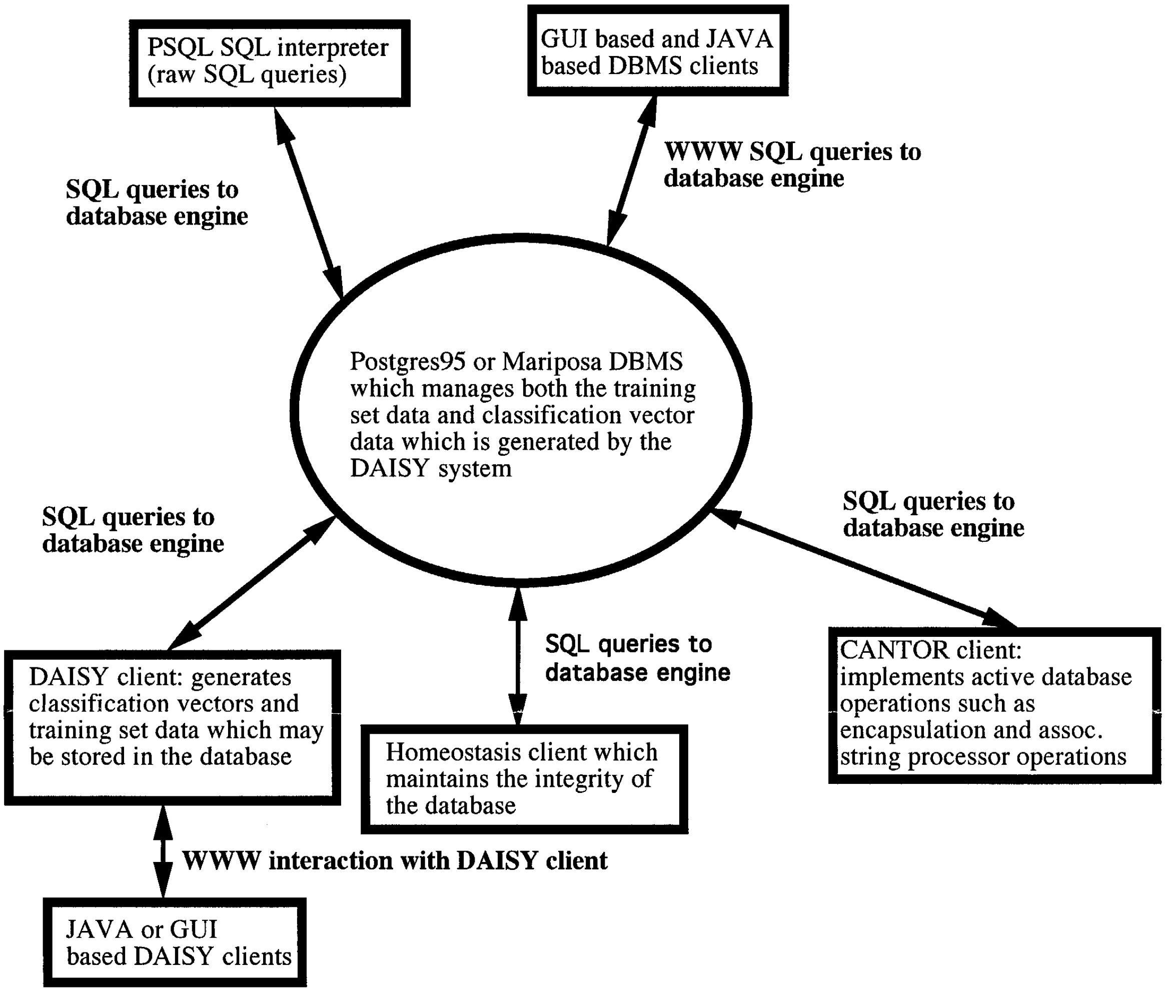
Figure 2. A schematic of the architecture of the DAISY classifier client.



Unknown image pathway
Autotraining data pathway
Results pathway
User supplied training data pathway
To Postgres95/Mariposa SQL DBMS
DAISY PCA associative memory classification engine
Postgres95/ Mariposa DBMS gateway
Internal cache of training set images
Sensitive directory interface
GUI interface
PSRP/JAVA interface
Sensitive training image directory
GUI based forms interface
PSRP client interface/JAVA applets
Sensitive unknown image directory
CCD camera
Khoros 2.2 GUI based image processing system (used for data capture and pre-processing)
Processed image files

(Ivins and Porrill, 1993; Schabel and Arridge, 1995) or caterpillars to automatically extract arbitrary regions of interest from the captured imagery; the use of image moment and/or envelope matching techniques to normalise wing area and orientation prior to training classifiers or identifying an unknown; the use of correlation vector metaclassification techniques; the use of aggregate (multi-character) classifiers; the use of higher level (genus) classifiers; the use of standard (SQL) database technology such as PostGres 95 (Yu and Chen, 1995) to keep track of data and training sets within the DAISY system; use of advanced adaptive correlation measures to improve species descrimination in the case of closely related sibling species. These advanced correlation measures will have to address two problems: (a) the effects of random correlation, which become increasingly important as image size is reduced; and, (b) the effects of pixel architecture. Effectively this means that adaptive ways of optimally weighting pixels in the **X**/**X**' image comparison must be sought. It is likely that this will be achieved by using genetic algorithm driven jack-knife training in conjunction with either pseudo random pixel weighting or local feature analysis (Penev and Attick, 1996); provision of World Wide Web access via Java enabled browses (e.g. NetScape); the application of DAISY PCA to wing areas previously

Figure 3. A schematic showing the interface between the DAISY system and SQL database engine (Postgres95/Mariposa DBMS).


PSQL SQL interpreter (raw SQL queries)
GUI based and JAVA based DBMS clients
SQL queries to database engine
WWW SQL queries to database engine
Postgres95 or Mariposa DBMS which manages both the training set data and classification vector data which is generated by the DAISY system
SQL queries to database engine
SQL queries to database engine
SQL queries to database engine
DAISY client: generates classification vectors and training set data which may be stored in the database
Homeostasis client which maintains the integrity of the database
CANTOR client: implements active database operations such as encapsulation and assoc. string processor operations
WWW interaction with DAISY client
JAVA or GUI based DAISY clients

identified by experts as containing discriminant characters; and application of DAISY algorithm to sub-areas of objects previously identified by experts as containing discriminant characters[4] .

The existing DAISY PCA classifier will be extended to enable it to deal effectively with three-dimensional objects. There are two ways of achieving this: (a) implementing a non-linear variant of the DAISY classifier; or, (b) linearisation of the three-dimensional data using optical stereo to generate *orthoimages* of the three-dimensional objects. Appropriate algorithms for accomplishing this (e.g. Lane *et al*., 1993; O'Neill and Denos, 1996) already exist. In addition, there is also a need to investigate the accuracy of the system using large (100 species +) wing databases, and to see whether the system can be used to identify specimens using other body parts. Block schematics of the proposed DAISY exemplar system to be developed under the aegis of Darwin Initiative funding is shown in Figs 2 and 3.

8. Conclusion

Computer aided identification systems of various forms have been explored on numerous occasions in the recent past. They have usually been rejected, explicitly or implicitly as an inappropriate tool for the task of making routine identifications. We suggest however that the results presented here, and in greater detail in the papers by Weeks *et al* (1987; submitted) support the claim that recent developments in the field of computer vision make the revisitation of this potential solution to the "taxonomic impediment" a high priority.

9. Acknowledgements

We thank Sondra Ward for preparing the wing slides and Jon Boorman for his advice about ceratopogonid taxonomy. This programme has been supported in part by AFRC grant 40/A01764, and subsequently under the aegis of Darwin Initiative funding. Kevin J. Gaston is a Royal Society University Research Fellow.

10. References

Ashbrook, A.P., Thacker, N.A., Rockett, P.I. and Brown, C.I. (1995) Robust recognition of scaled shapes using pairwise geometric histograms. *Proceedings of the British Machine Vision Conference*, 503-512.

Gauld, I.D. (1986) Taxonomy, its limitations and its role in understanding parasitoid biology. In: Waage J. and Greathead D. (eds), *Insect Parasitoids*. Academic Press, London, pp 1-21.

Ivins, J. and Porrill, J. (1993) Active region models for segmenting colours and textures. *Image and Vision Computing* **13**, 431-438.

Lane, R.A. Thacker, N.A. and Seed, N.L. (1993) Stretch correlation as an alternative to feature based stereo matching algorithms. *Image and Vision Computing* **12**, 203-212.

O'Neill, M.A. and Denos, M.I. (1996) Automated system for coarse-to-fine pyramidal area correlation stereo matching. *Image and Vision Computing* **14**, 225-236.

O'Neill, M.A and Hilgetag, C.C. (1998) *CANTOR – A novel system for the analysis and classification of complex data*. In prep.

Pankhurst, R.J. (1978) *Biological Identification*. Arnold, London.

[4] This effectively permits DAISY classifiers to be embedded within expert systems and interactive keys.

Penev, P.S. and Attick, J.J. (1996) Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems* **7**, 477-500.

Pentland, A., Moghaddam, B. and Starner, T. (1994) View-based and modular eigenspaces for face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, 1994, 84-91.*

Schabel, J.A. and Arridge, S.R. (1995) Active contour models for shape description using multiscale differential invariants. *Proceedings of the British Machine Vision Conference, 1995,* 197-206.

Turk, M., and Pentland, A. (1991) Eigenfaces for recognition. *Journal of Cognitive Neurosciences* **3**, 71-86.

Valentin, D., Abdi, H., O'Toole, A.J. and Cottrell, G.W. (1994) Connectionist models of face processing: a survey. *Pattern Recognition* **27**, 1209-1230.

Weeks, P.J.D. and Gaston, K.J. (1997) Image analysis, neural networks and the taxonomic impediment to biodiversity. *Biodiversity and Conservation* **6**, 263-274.

Weeks, P.J.D., Gauld, I.D., Gaston, K.J. and O'Neill, M.A. (1997) Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research* **87**, 203-211.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D. [Submitted, 1997] Development of an automated insect identification system. *Journal of Applied Entomology*.

White, I.M. and Scott, P.R. (1994) Computer information resources for pest identification: a review. In: Hawksworth, D.L. (ed.), *The Identification and Characterisation of Pest Organisms*. CAB International, Wallingford, pp.129-137.

Young, M.P., Scannell, J.W., O'Neill, M.A., Hilgetag, C.C., Burns, G. and Blakemore, C. (1996) Non-metric multi-dimensional scaling in the analysis of neuroanatomical connection data and the organisation of the primate cortical visual system. *Proceedings of the Royal Society of London*, B **348**, 281-308.

Yu, A. and Chen, J. (1995) *The PostGres95 Users Manual.* University of California at Berkeley, Version 1.0.

Yu, D.S., Kokko, E.G., Barron, J.R., Schaalje, G.B. and Gowen, B.E. (1992) Identification of ichneumonid wasps using image analysis of wings. *Systematic Entomology* **17**, 389-395.

DEVELOPING IDENTIFICATION THRESHOLDS FOR AUTOMATED IDENTIFICATION

By P.D. Bridge
CABI Bioscience, Bakeham Lane, Egham, Surrey TW20 9TY

1. Introduction

There are a number of possible outcomes from any identification scheme, and the two most desirable are that an unknown organism may be identified correctly through the scheme, or may fail to identify. In some cases unknown organisms may be incorrectly identified, and depending on the type of identification scheme, both correct and incorrect identifications may be qualified or confirmed by numerical values (see Bridge, 1997). The actual number of outcomes available will depend on the type of identification scheme, however, the common point in all schemes is a decision that the unknown organism is identified. In the development of an automated system there needs to be some limit or threshold put onto this decision, so that an unknown is considered identified only when a particular criterion, or set of criteria, has been met.

2. Specific Features of Microbiological Data

In microbiology the characters available to describe and define taxa are often quantitative in terms of occurrence. That is, many taxa are defined by characters which are shown by most, but not always all, representatives of the group. This data can be presented in terms of a frequency table which gives the frequency of occurrence of each character within each taxon (see Sneath and Sokal, 1973; Willcox *et al.*, 1973; 1980). This type of data has been used in identification schemes with probability and distance measures, or in profile matching schemes (e.g. Willcox *et al.*, 1980; Priest & Alexander, 1988; Barnett *et al.*, 1990). More recently new techniques such as neural networks have been investigated (e.g. Morris *et al.*, 1992; Rataj and Schindler, 1991).

3. Existing Cutoffs and Schemes

Computer based and computer assisted identification schemes are well established in microbiology, and have been used in the routine identification of medically important bacteria for more than 20 years (e.g. Willcox *et al.*, 1973; Feltham and Sneath, 1982). Most established schemes are of one of two types, that is they are based on either a probability or a distance measure. Probabilities and distances coefficients are often mathematically equivalent, and both are dependent on the original reference data. As a result, if the original reference data contains a lot of variable results, the subsequent identification values will be low. Secondly, both probabilities and distance values are sensitive to the total number of characters contained in the scheme, and so some form of scaling is generally required. Thirdly, probabilities and distances are sensitive to the total number of taxa in the scheme, and some form of normalisation or comparison is generally required. As a result of these factors, the threshold level for an acceptable identification will vary depending on the database used for the identification. Therefore if an automated system is to be developed, then the identification threshold will need to be considered for each database used. For example, in a recent comparison of two fungal identification schemes cutoff levels of about 0.85 were found to be sufficient for good identification with the Willcox probability measure (Bridge, 1997). Other studies had however recommended values from 0.85 to 0.99 with other data sets (see Table 1). While each different value proved acceptable for the data being analysed, none could be taken as a general cutoff.

Table 1. Some Critical Probabilistic Values from Bacteriology and Mycology

	Cutoff score	**Acceptable score**	**Reference**
Gram negative bacteria	0.999	0.95	Willcox *et al.*, 1973
Gram positive bacteria	0.999	0.9-0.95	Feltham and Sneath, 1982
Streptomycetes	0.85	0.8	Williams *et al.*, 1983
Yeasts*	approx 0.85	0.7	Bridge, 1997
*Penicillium**	0.85	0.65	Bridge *et al.*, 1992

*Additional coefficients or measures included

The decrease in acceptable identification score does not always show a simple relationship to the number of discrepant characters in the identification. One reason for this is that the relationship will be complicated by the actual frequency values of the individual characters, the closeness of other taxa and the total number of characters included in the scheme (see Bridge, 1997). Away from probability measures, distance measures can suffer in similar ways, the critical distance from the centre of a group being dependent on the size of the original group, and the closeness of neighbouring groups (see Sneath and Sokal, 1973; Gyllenberg and Niemellä, 1975; Sneath, 1979a).

4. Database Construction and Selection of Features

An important consideration in the development of an automated identification system is the level of identification required, not only as a threshold value, but also as a practical outcome. For most identification systems there may be additional confirmatory characteristics which are not used in the identification system, but which are used in a more traditional manner. For example, if trying to identify a bacterium from a clinical sample many plant-pathogenic species are of little relevance. These species may be closely related to the human pathogen, but never occur in that environment. Therefore they may be discounted from the identification process without the need to calculate scores or matches. It needs to be remembered that any identification scheme is being constructed for a practical purpose, and so should be appropriate to the requirements specified by the "end-user". One example of where this discounting has already been performed in the construction of the original data-base, is in systems which are designed to identify organisms from a particular niche or habitat, e.g. an identification scheme for clinically associated yeasts (bioMérieux, 1993). However, this is not always the case and more systematic databases, such as for example a scheme to identify all members of a genus, may not include such information.

A similar situation may exist with particular features within a database, where individual characters do not have equal weights. In such cases while it is acceptable for some characters to be variable, and therefore for mismatches to be allowed, some other characters will be definitive for the species or genus, and a mismatch in these characters will always lead to a failure to identify. Most numerical systems have the facility to consider this situation, such as through the inclusion of prior probabilities (see Willcox *et al.*, 1980), but again these values may vary considerably between different taxa and so may not always provide a practical solution. Characters that are of crucial importance in the identification of some taxa may be of negligible importance for others in the same database. This requires a separate prior probability to be known for every character for every taxon, even then the differing levels of weighting may not be accurately conveyed within the final calculation. In Bayesian systems there is a tendency for prior probabilities to cancel out, and their effect can be lost in a large scheme. A further problem with prior probabilities is that in order to be entirely accurate they need to be based on all members of the taxon, and in most cases this information is not available.

5. Test Selection

Following on from the last point, the selection of tests for the identification system is one of the key elements in the construction of the final scheme. This is a procedure that can be automated and there are numerous routines which enable the calculation of the information content of a particular character, or the measurement of some level of identification potential (see Sneath, 1979b; Bridge *et al*., 1992). These enable the ranking of characters and can lead to the selection of the most appropriate character at any point in a scheme. An important consideration in the use of these techniques is what type of character is required for the identification scheme. Some characters may appear suitable for an identification scheme in that there are a large proportion of positive and negative results and the test effectively splits all the taxa into two roughly equal sized groups. A series of such tests would produce a fairly balanced dichotomous key. Alternatively, some characters may show a single response for a large number of taxa, while separating out perhaps one or two. This type of character would be particularly useful as diagnostic or confirmatory character in the latter stages of an identification (see Bridge *et al*., 1992). Alternatively diagnostic characters can be used at the beginning of an identification scheme to remove outliers such as very different taxa. This may not be necessary in an identification scheme for a single taxonomic group, but may be very important in a scheme based on very different taxa from a common environment or condition.

Which character type is preferred will be very dependent upon the user and the particular circumstances associated with the scheme. It is important to remember that some of the routines used to select tests will tend to suggest mainly “key”-like characters, such as the separation potential measures (Sneath, 1979b), whereas other routines will tend to select more “diagnostic” characters, such as Gyllenberg’s Sum of C (Gyllenberg, 1963).

6. An Overall System

Although there are obvious problems with developing automated cutoff values, their function is crucial for the success of any identification scheme. There are some general trends that can be identified from numerical methods, and combining these with an expert specific knowledge of the identification data can bring about sets of criteria that will be suitable for arriving at judgements with specific data sets. This approach has been considered previously by a variety of authors (e.g. Sneath, 1979a; D’Amato *et al*., 1981; Holmes and Hill, 1985), and the following scheme is based upon findings from these studies and two fungal identification schemes (Bridge *et al*., 1992; Bridge, 1997).

Firstly, there should be a low proportion of tests against the identification, this will be very dependent on what is considered mismatch, and the proportion of very variable characters for the taxon being considered. Secondly, of the commercially used and accepted numerical identification software, two of the most commonly used coefficients are a normalised probability score and some form of modal likelihood or fractional distance (see D’Amato *et al*., 1981; bioMérieux, 1993; Bridge, 1997). Although the modal likelihood is called a likelihood coefficient it is mathematically equivalent to a distance value in that it provides a score that is directly related to the taxon centroid (Dybowski and Franklin, 1968). Distance scores are open ended, in that they will continue to get larger as the unknown becomes less likely. While normalised probabilities tend to be very high for good identifications, when the identification is poor or incorrect the scores tend to fall away quickly. Therefore one identification threshold that can be set in numerical systems is for a high normalised score, few test against identification and a low distance value (Sneath, 1979a). Variations in either coefficient may be indicative of different occurrences. For example, a low probability and a low distance value generally implies that the reference groups are very close together, and so a clear delineation between them may not be possible (Bridge, 1997). In another case it is possible to get a high normalised score for an incorrect identification. This can occur when although the probability of the unknown belonging to one group is very low, it is so much lower for all other groups that normalising gives a high value (Sneath, 1979a; Bridge *et al*., 1992). In this case it is often due to the unknown not being included in the reference data and so the distance score is very large. A good probability score and a slightly larger than average distance measure may be obtained when the unknown is a member of a well separated but very variable group.

The combination of the two identification measures together can give a framework for an automated threshold system (see Table 2), but the cut-off values that will be used will be dependent on the database. Unfortunately there would appear to be no clear method for determining *a priori* scores for the probability and distance measures. It is in all cases desirable to undertake a reasonable level of testing of any new identification scheme, and this process can be used to determine at least initial cut-off values. The most satisfactory so far available is repeated testing of specimens of known identity, in order to see the range of identification scores and criteria obtained. This is not always possible, even with microbiological data, where there is often a shortage of verified isolates for some taxa. One possibility is to generate artificial typical and atypical character sets (see Sneath, 1980; Williams *et al*., 1985; Bridge *et al*., 1992). One problem with this is that the generation of artificial character sets needs to be undertaken within the bounds of the known variability of the taxa, and so again becomes very database specific. A temporary solution that allows at least some confidence to be placed in the scheme, and provides a realistic starting point for setting cutoff levels is the re-identification of the constituent taxa from the database. This is in many ways similar to the supervised learning of a neural network (Boddy *et al*., 1990), and analogous to boot-strapping.

A final method of verification that can also add confidence to an identification is to include within the system some of the expert knowledge that may be used by the taxonomist in the final interpretation of a result. This can be in many forms, such as the inclusion of prior probabilities, or can be a simple verification check based on the characters considered most important for that taxon.

7. Conclusion

In conclusion, the specific features of any identification data set will affect the level at which identifications can be made from it. As a result it is generally not possible to recommend universal automatic cutoff levels for numerical identification scores. There are however a number of common frameworks into which cutoff levels can be included, and suitable cutoffs for individual data sets can be generated by repeated testing of both real and artificially generated data sets. Repeated testing is a feature of some numerical and neural network schemes and is at present the most reliable starting point for the evaluation of any scheme.

Table 2. Potential Framework for Threshold Criteria for Probability and Distance Based Schemes

Criteria
1. Low proportion of tests against ID: This can be measured for the scheme as a whole, or for individual taxa. "Bad" taxa may be masked by very good ones in total scheme comparisons.
2. High probability score: Should exceed a minimum cutoff value to give an identification, but should be considered in relation to other criteria.
3. Low distance score: Cutoff values difficult to determine when taxa contain different levels of variability, can be scaled according to most typical organisms or according to variability in group.
Some possible decisions from combined criteria: Low tests against, high probability, low distance - good ID. Low tests against, high probability, high distance - ID to very variable group? High tests against, high probability, high distance - unknown not in database? High tests against, moderate probability, moderate distance - atypical member of group? Low tests against, low probability, low distance score - overlapping taxa.

8. References

Barnett, J.A., Payne, R.W. and Yarrow, D. (1990) *Yeast Identification PC Program*. Barnett, Norwich.

bioMérieux (1993) *ID32C Analytical Profile Index*. BioMérieux, Marcy-l'Etoile.

Boddy, L., Morris, C.W. and Wimpenny, J.W.T. (1990) Introduction to neural networks. *Binary* **5**, 17-22.

Bridge, P.D. (1997) Mixing elements from different identification systems. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 233-245.

Bridge, P.D., Kozakiewicz, Z. and Paterson, R.R.M. (1992) PENIMAT: A computer assisted identification scheme for terverticillate *Penicillium* isolates. *Mycological Papers* **165**, 1-59.

D'Amato, R.F., Holmes, B. and Bottone, E.J. (1981) The systems approach to diagnostic microbiology. *CRC Critical Reviews in Microbiology* **9**, 1-44.

Dybowski, W. and Franklin, D.A. (1968) Conditional probability and identification of bacteria: a pilot study. *Journal of General Microbiology* **54**, 215-229.

Feltham, R.K.A. and Sneath, P.H.A. (1982) Construction of matrices for computer-assisted identification of aerobic Gram-positive cocci. *Journal of General Microbiology* **128**, 713-720.

Gyllenberg, H.G. (1963) A general method for deriving determination schemes for random collections of microbial isolates. *Annals of the Academy of Sciences Fenn. Series A. IV Biology, No. 69.*

Gyllenberg, H.G. and Niemelä, T.K. (1975) New approaches to automatic identification of micro-organisms. In: Pankhurst, R.J. (ed.) *Biological Identification with Computers*. Systematics Association Special Volume No.7. Academic Press, London.

Holmes, B. and Hill, L.R. (1985) Computers in diagnostic bacteriology, including identification. In: Goodfellow, M., Jones, D. and Priest, F.G. (eds), *Computer-assisted Bacterial Systematics*. Academic Press, London, pp 265-288.

Morris, C.W., Boddy, L. and Allman, R. (1992) Identification of basidiomycete spores by neural network analysis of flow cytometry data. *Mycological Research* **96**, 697-701.

Priest, F.G. and Alexander, B. (1988) A frequency matrix for probabilistic identification of some bacilli. *Journal of General Microbiology* **134**, 3011-3018.

Rataj, T. and Schindler, J. (1991) Identification of bacteria by multilayer neural network. *Binary* **5**, 9-12.

Sneath, P.H.A. (1979a) BASIC program for identification of an unknown with presence-absence data against an identification matrix of percent positive characters. *Computers and Geosciences* **5**, 195-213.

Sneath, P.H.A. (1979b) BASIC program for character separation indices from an identification matrix of percent positive characters. *Computers and Geosciences* **5**, 349-357.

Sneath, P.H.A. (1980) BASIC program for determining the best identification scores possible from the most typical examples when compared with an identification matrix of percent positive characters. *Computers and Geosciences* **6**, 27-34.

Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*. W.H. Freeman, San Francisco.

Willcox, W.R., Lapage, S.P., Bascomb, S. and Curtis, M.A. (1973) Identification of bacteria by computer: theory and programming. *Journal of General Microbiology* **77**, 317-330.

Willcox, W.R., Lapage, S.P. and Holmes, B. (1980) A review of numerical methods in bacterial identification. *Antonie van Leeuwenhoek* **46**, 233-299.

Williams, S.T., Goodfellow, M., Wellington, E.M.H., Vickers, J.C., Alderson, G., Sneath, P.H.A., Sackin, M.J. and Mortimer, A.M. (1983) A probability matrix for identification of some streptomycetes. *Journal of General Microbiology* **129**, 1815-1830.

Williams, S.T., Vickers, J.C. and Goodfellow, M. (1985) Application of new theoretical concepts to the identification of streptomycetes. In: Goodfellow, M., Jones, D. and Priest, F.G. (eds), *Computer-assisted Bacterial Systematics*. Academic Press, London, pp 289-306.

ARTIFICIAL NEURAL NETWORKS FOR IDENTIFICATION

By L. Boddy
School of Pure and Applied Biology, University of Wales, Cardiff CF1 3TL

C. W. Morris
Department of Computer Studies, University of Glamorgan, Pontypridd CF37 1DL

1. Introduction

Traditionally biological identification has been carried out either by biologists that are extremely knowledgeable about the taxa which are to be identified or by skilled biologists using keys to aid the identification process. Unfortunately, such experts are now becoming rare and there is great difficulty in obtaining rapid, accurate identifications not least in developing countries where the cost of using such experts may be prohibitive. A partial solution to this problem would be to develop automated identification systems employing computer technology. Some attempts have been made in this direction, for example, PANKEY was a first attempt to computerise the more traditional paper-based key. CABIKEY is a more sophisticated product which uses multimedia techniques to aid the identification process (White and Scott, 1994; White, this volume). ROTTERS takes a slightly different approach being an expert system to identify some wood decay fungi (Rose, 1993). Whilst all these are significant improvements over the traditional key-based approach they still require considerable biological skill from the user if they are to be employed with any success.

Recent development of high technology equipment provides opportunities to implement fully automated identification systems that require little or no biological expertise. Such technologies include image analysis (e.g. DAISY; Weeks *et al*., 1997), acoustic recognition (see Chesmore, this volume), pyrolysis mass spectrometry (see references in Table 2), molecular techniques and flow cytometry (see below). These systems all measure different features which are then used to distinguish taxa. Analysing the data is, however, often difficult. Typically there are large quantities of high dimensional data whose underlying distribution is unknown, which poses problems for traditional statistical techniques, thus more advanced non-parametric statistics or artificial neural networks (ANNs) may need to be applied.

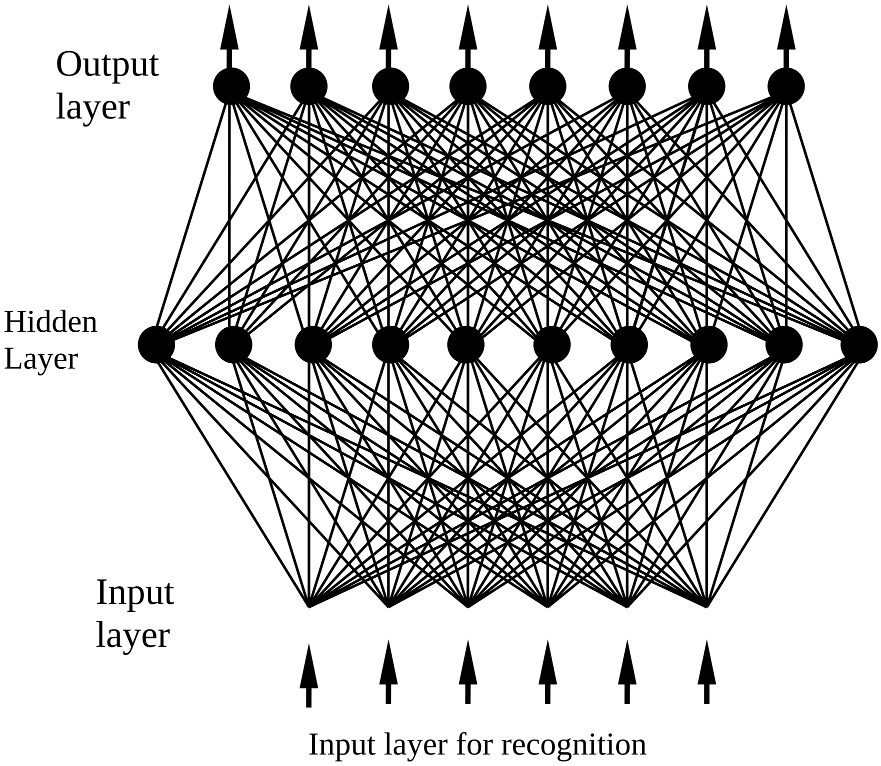
This paper will discuss ANNs as a suitable data analysis tool and then use flow cytometry as an example of an automated identification system.

2. Artificial Neural Networks

ANNs have been developed to model the operation of the brain in a limited way in terms of its ability to learn and discriminate. There are many different types of ANN but these can all be separated into two distinct classes: supervised and unsupervised networks. These two network types perform different tasks with supervised networks acting as identification systems and unsupervised networks able to perform classification.

Supervised networks require examples of the taxa that are to be identified. These examples must be in the form of numeric features that are used as the input to the network. Note that qualitative information can also be used (with care) but will need to be encoded into a numerical form. The network also requires information relating to the correct identity of each of the examples that are being used. The network will typically be organised as shown in Figure 1 with an input layer (one input per feature), a hidden layer (optimum size obtained by experiment) and an output layer (one output per taxon to be identified). The network will take each example and will learn to distinguish between the different taxa upon which it is being trained. This training is carried out automatically with the network altering its internal organisation so that it can successfully distinguish between all the examples presented. Once the network has learnt all of the training examples its performance must be assessed using further examples upon which the network has not been trained. This independent test set gives a measure of the network performance and should be used to gauge the usefulness (i.e. ability to generalise) of the net rather than the ability of the net to learn the training set. Typical examples of supervised ANNs are back propagation and radial basis function networks.

Figure 1. Feedforward architecture common to many ANNs used for identification problems. There is one input node per character, and one output node per taxon upon which the network was trained.



The training phase is an iterative process and can be very time consuming. However once trained the ANN can function fast enough to work in real time for many applications. The ANN will have formed its own relationships between the input characters and will thus be able to generalise about the examples that it has learnt. Once trained an ANN can cope remarkably well with data that are fuzzy, incomplete or partially contradictory. The major caveat here is that the network needs to be trained on a representative sample of the classes which it is to identify. As with all data analysis techniques it can only perform as well as the data allows it to.

3. Case Study: Flow Cytometry for Identification and Quantification of Phytoplankton Species

Traditionally, in order to investigate the species of phytoplankton present in a sample of water it has been necessary to perform microscopic identification which requires highly skilled personnel. Also, this approach is very slow and samples taken can only be analysed at a later date which means that interesting findings cannot be further investigated immediately. Flow cytometry may provide a solution being a technique that allows rapid, simultaneous measurements to be made on single particles (Burkill and Mantoura, 1990; Boddy and Morris, 1993; Jonker *et al*., 1995). A focused stream of particles is illuminated by a laser (or lasers) and measurements are made on the effect of the particle on the laser beam. Typically about 4-12 measurements are made on each particle and these include time of flight through the laser beam (an indication of particle size), various fluorescences (which are used to detect chemical content, e.g. photosynthetic pigments), scatter and polarisation (which give some idea of structure). For each particle these measurements are logged hopefully to give a distinguishing fingerprint. The system will measure of the order of 10^3 cells per second thus creating huge data sets very quickly.

The analysis of these large data sets poses a major problem. Scatter plots of pairs of measurements can allow skilled observers to identify clusters of cells but clearly this is not the most appropriate technique. Multivariate statistics have been used but these techniques have been shown to be time consuming to implement for the size of data sets produced (e.g. Demers *et al*., 1992). Neural networks, however, can cope with these data and are able to analyse the data in near real time (e.g. Boddy and Morris, 1993; Boddy *et al*., 1994b; Morris *et al*., 1994; Wilkins *et al*., 1994a,b, 1996).

ANNs have been trained on flow cytometry data obtained from pure cultures of up to 40 common phytoplankton species. The resultant networks were tested using unseen data sets and misclassification matrices constructed (Table 1). The latter indicates along the leading diagonal the proportion of test data patterns assigned the correct identity, while the values in the rest of the matrix indicates what each taxon was misidentified as. Misidentifications result from overlap of character distributions between taxa, and are not a fault of the ANN. They can only be resolved by using additional characters. Confidence of classification (i.e. frequency of correct classification divided by total frequency that a taxon was assigned an identity) is clearly very important. For example, a network that identifies species X correctly 100% of the time appears to have good performance. However, if that same network identifies species Y as species X 50% of the time, when the network makes an identification as X there is uncertainty about the true identity of the particle.

An even bigger problem arises when the system is presented with a taxon upon which it has not been trained. The ANN (or indeed any data analysis tool) needs to be able to reject as unknown such individuals and not simply match them to the nearest known taxon. Radial basis function ANNs have the capability of performing this rejection making them prime candidates for the data analysis part of any automated identification tool (Morris and Boddy, 1996).

4. Conclusions

The quest for automated identification systems will almost certainly lead to vast quantities of data to be analysed. These data will take many forms but will be characterised by being multidimensional and often non-normal. The challenge for the designers of these automated systems will be to develop high performance data analysis techniques. Neural networks appear to offer a powerful solution to this problem. Table 2 shows the current state of the art in the application of neural networks to automated identification. It can be seen from this table that most applications have only looked at identification of few species, or where more have been examined, they may in some cases, be easily distinguishable by other techniques anyway. If these techniques are to be used for realistic problems then they need to be scaled up in terms of the number of taxa to be distinguished, and we are currently investigating techniques for this.

Table 1. Misclassification matrix for a 48 hidden node Radial Basis Function network trained using 400 different patterns per taxon, with 11 inputs (light scatter, diffraction and fluorescence parameters) trained over two complete presentations of the training data set and tested using ‘unseen’ test data. The 12 different species of marine and freshwater phytoplankton were grown under controlled conditions to give pure cultures and analysed using the EurOPA flow cytometer (Jonker *et al*., 1995).

Species name	**No.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Porphyridium purpureum	1	**.99**	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00
Dunaliella tertiolecta	2	.00	**.98**	.00	.00	.00	.00	.00	.03	.00	.00	.00	.00
Chlorella salina	3	.00	.00	**.85**	.07	.00	.01	.03	.00	.05	.00	.00	.00
Stichococcus bacillaris	4	.00	.00	.02	**.74**	.05	.04	.05	.00	.02	.03	.06	.00
Ochromonas sp.	5	.00	.00	.01	.08	**.87**	.01	.01	.00	.00	.01	.01	.01
Pseudopedinella sp. 1	6	.01	.00	.02	.02	.01	**.73**	.03	.00	.00	.05	.11	.02
Pseudopedinella sp. 2	7	.00	.00	.08	.01	.00	.00	**.88**	.00	.00	.00	.03	.00
Pyramimonas obovata	8	.00	.05	.00	.00	.00	.01	.00	**.94**	.00	.00	.00	.00
Halosphaera russellii	9	.00	.01	.02	.00	.00	.01	.05	.00	**.90**	.00	.00	.02
Nephroselmis rotunda	10	.00	.00	.00	.01	.00	.02	.00	.00	.00	**.97**	.00	.00
Pyramimonas grossii	11	.00	.00	.09	.11	.01	.11	.03	.00	.00	.01	**.65**	.01
Tetraselmis verrucosa	12	.00	.00	.01	.00	.00	.01	.00	.00	.01	.00	.01	**.96**
Confidence of classification		**.98**	**.95**	**.79**	**.72**	**.93**	**.77**	**.83**	**.97**	**.92**	**.91**	**.74**	**.94**

That ANNs have a great deal to offer is further emphasised by comparison with more traditional key-based techniques, probabilistic and expert systems (Table 3). If ANNs are being used with

Table 2. Some Recent Uses of ANNs in Identification and Clustering.

No. of taxa to be discriminated	No. of taxa succesfully (>80% of replicates) discriminated	Organisms	Measurement techniques	Authors	Comments
4 species	all	fungal spores	flow cytometry	Morris, Boddy and Allman (1992)	backpropagation trained on 50 spores per species
3 plus 2 sizes of beads	all	microalgae & beads	flow cytometry	Frankel *et al.* (1989)	backpropagation and Kohonen ANN trained on > 4500 cells
8 groupings	all	microalgae	flow cytometry	Balfoort *et al.* (1992); Smits *et al*. (1992)	backpropagation trained on 500, 1000 and 2000 cells
6 groupings	all	cyanobacteria	flow cytometry	Frankel *et al*. (1996)	backpropagation trained on a total of 6000 cells
5 species	all	*Penicillium* spp.	cultural characters	Bridge, Morris and Boddy (1994)	backpropagation and Kohonen ANNs trained on 1 strain per taxon, 12 inputs
3 species plus a group of ‘others’	all	*Streptomyces* spp.	pyrolysis mass spectrometry	Chen *et al*. (1993a,b)	backpropagation trained on a total of 27 pyrolysis mass spectra
2 ‘species’	all	*Mycobacterium tuberculosis* & *M. bovis*	pyrolysis mass spectrometry	Freeman *et al*. (1994)	backpropagation trained on 8 strains of *M.b.* & 13 strains of *M.t.*
3 ‘species’ (11 strains) plus a group of ‘others’	all	*Propionibacterium* spp.	pyrolysis mass spectrometry	Goodacre *et al*. (1994)	backpropagation; trained on 4 pyrolysis mass spectra; Kohonen SOM
35 strains (9 ‘species’)	all	oral *Eubacterium* strains & abscess *Peptostreptococcus* sp.	pyrolysis mass spectrometry	Goodacre *et al*. (1996)	backpropagation trained on 3 pyrolysis mass spectra per strain
11 species	all	fungi which invade standing trees	cultural charcaters	Morgan (1997)	RBF ANNs trained on 25 strains per taxon
16 - 24 species	88%	Fungal (*Pestalotiopsis* & related species) spores	morphometric data	Boddy *et al*. (1994a), Morgan *et al*. (1998)	backpropagation, RBF, ARBF, LVQ, Kohonen ANNs trained on 25 spores per taxon
12 - 40 species	variable* 50 - 85%	microalgae	flow cytometry	Boddy *et al*. (1994b), Morris *et al.* (1994), Wilkins *et al.* (1994a,b; 1996), Morris & Boddy (1995)	backpropagation, RBF, ARBF, LVQ, Kohonen ANNs trained on 200 - 400 cells per taxon
101 ‘species’	96%	Enteric bacteria (Gram negative rods)	possibly cultural/ biochemical chars	Schindler, Farmer and Paryzek (1992)	backpropagation trained on a total of 3429 strains, 47 inputs

* Note that where successful identification was relatively low, this reflects overlap of distributions of characters used for discrimination, rather than inadequacies of the ANN technique.

Table 3. A Comparison of Attributes of Different Computer Identification Approaches. Based Partly on Woolley and Stone (1987) and Morris and Boddy (1995).

Attribute	Dichotomous key	Tabular keys / profile matching	Multiple entry key / polyclave	Probability and distance methods	Expert systems	Artificial neural networks
Structural efficiency	Some	no	user[2]	some	yes	Hierarchical ANNs can be made so
Dynamic efficiency	no	no	user[2]	no	yes	Hierarchical ANNs can to some extent
Missing data[1]	no	yes	yes	yes	yes	yes
Incorrect data[1]	no	limited	limited	yes	limited	yes
Indication of absence from database	no	no	to some extent	to some extent	yes	RBF and ARBF networks can
Indicating possibilities if more than one	no	yes	yes	yes	yes	yes
Explicit probabilities	no	no	no	yes	yes	yes
Easy updating	no	yes	yes	no	some	Most require complete retraining
Transparent logic	yes	N/A	user	N/A	yes	ANNs do not employ logic but learn; methods of ‘interrogating' them are becoming available
Domain knowledge required for construction	yes	yes	yes	yes	yes	valuable but not essential

[1] Depends on which characters are missing and their importance as major discriminators.
[2] The user can be directed to observe characters which will lead to an efficient identification.

systems where data collection is not automated a disadvantage is that they do not possess structural or dynamic efficiency, which means that a lot of time might be wasted in recording characters of a specimen that may not assist in making the identification. Another slight disadvantage is that with many types of ANN the network has to be completely retrained when additional taxa are to be incorporated. However, adopting an hierarchical approach can lessen the impact of both of these problems. Some taxonomists may consider that lack of transparent logic (i.e. the ability to determine directly how/why a particular identification was made) is a major drawback, however, methods are being developed to interrogate ANNs.

On balance ANNs probably have more in favour of their use than other methods but they should not be expected to be the best solution to all problems. There is no universal solution.

5. References

Balfoort, H.W., Snoek, J., Smits, J.R.M., Breedveld, L.W., Hofstraat, J.W. and Ringelberg, J. (1992) Automatic identification of algae: neural network analysis of flow cytometric data. *Journal of Plankton Research* **14**, 575-589.

Boddy, L., Gimblett, A.M., Morris, C.W. and Mordue, E.J.M. (1994a) Neural network analysis of fungal spore morphometric data for identification of species in the genus *Pestalotiopsis*. In: Dagli, C.H., Fernandez, B.R., Ghosh, J. and Kumara, R.T.S. (eds), *Intelligent Engineering Systems Through Artificial Neural Networks,* Vol. **4**. ASME Press, New York, pp. 605–612.

Boddy, L. and Morris, C.W. (1993) Analysis of flow cytometry data: a neural network approach. *Binary* **5**, 17-22.

Boddy, L., Morris, C.W., Wilkins, M.F., Tarran, G.A. and Burkill, P.H. (1994b) Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry* **15**, 283-293.

Bridge, P.D., Boddy, L. and Morris, C.W. (1994) Information resources for pest identification - an overview of computer-aided approaches. In: Hawksworth, D.L. (ed.), *The Identification and Characterisation of Pest Organisms*. CAB International, Wallingford, pp. 153-167.

Burkill, P.H. and Mantoura, R.F.C. (1990) The rapid analysis of single marine cells by flow cytometry. *Philosophical Transactions of the Royal Society* **A 333**, 99-112.

Chen, J., Atalan, E., Ward, A.C. and Goodfellow, M. (1993a) Artificial neural network analysis of pyrolysis mass spectrometric data in the identification of *Streptomyces* strains. *FEMS Microbiology Letters* **107**, 321-326.

Chen, J., Atalan, E., Kim, H.J., Hamid, M.E., Tru-Jillo, M.E., Magee, J.G., Mafio, G., Ward, A.C. and Goodfellow, M. (1993b) Rapid identification of streptomycetes by artificial neural network analysis of pyrolysis mass spectra. *FEMS Microbiology Letters* **114**, 115-120.

Demers, S., Kim, J., Legendre, P. and Legendre, L. (1992) Analysing multivariate flow cytometric data in aquatic sciences. *Cytometry* **13**, 291-298.

Frankel, D.S., Olson, R.J., Frankel, S.L. and Chisholm, S.W. (1989) Use of a neural net computer system for analysis of flow cytometric data of phytoplankton populations. *Cytometry* **10**, 540-550.

Frankel, D.S., Frankel, S.L., Binder, B.J. and Vogt, R.F. (1996) Application of neural networks to flow cytometry data analysis and real-time cell classification. *Cytometry* **23**, 290-302.

Freeman, R., Goodacre, R., Sisson, P.R., Magee, J.G., Ward, A.C. and Lightfoot, N.F. (1994) Rapid identification of species within the Mycobacterium tuberculosis complex by artificial neural network analysis of pyrolysis mass spectra. *Journal of Medical Microbiology* **40**, 170-173.

Goodacre, R., Neal, M.R., Kell, D.B., Greenham, L.W., Noble, W.C. and Harvey, R.G. (1994) Rapid identification using pyrolysis mass spectrometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs. *Journal of Applied Bacteriology* **13**, 157-160.

Goodacre, R., Hiom, S.J., Cheeseman, S.L., Murdoch, D., Weightman, A.J. and Wade, W.G. (1996) Identification and discrimination of oral asaccharolytic *Eubacterium* spp. by pyrolysis mass spectrometry and artificial neural networks. *Current Microbiology* **32**, 77-84.

Jonker, R.R., Ringelberg, J., Dubelaar, G.B.J., Konig, J.W., Van Veen, J.J.F., Wietzorrek, J., Kachel, V., Cunningham, A., Burkill, P.H., Tarran, G., Wilkins, M.F., Boddy, L., Morris, C.W. and Peeters, J.C.H. (1995) A European Optical Plankton Analysis System: flow cytometer based technology for automated phytoplankton identification and quantification. In: Weydert, M., Lipiatou, E., Goni, R., Frangakis, C., Bohle-Carbonell, M. and Barthel, K.-G. (eds) *Marine Science and Technologies. 2nd MAST Days and EUROMAR Market*. CEC, Brussels, pp. 945-995.

Morgan, A. (1997) *The Application of Artificial Neural Networks to Fungal Taxonomy and Identification.* Ph.D. Thesis, University of Wales, Cardiff.

Morgan, A., Boddy, L., Mordue, J.E.M. and Morris, C.W. (1998) Identification of species in the genus *Pestalotiopsis* from spore morphometric data: a comparison of some neural and non-neural methods. *Mycological Research* **102**, 975-984.

Morris, C.W. and Boddy, L. (1995) Artificial neural networks in identification and systematics of eukaryotic microorganisms. *Binary* **7**, 70-76.

Morris, C.W. and Boddy, L. (1996) Classification as unknown by RBF networks: discriminating phytoplankton taxa from flow cytometry data. In: Dagli, C.H., Akay, M.C., Chen, L.P., Fernandez, B.R., Ghosh, J. and Kumara, R.T.S. (eds), *Intelligent Engineering Systems Through Artificial Neural Networks,* Vol. **6**. ASME Press, New York, pp. 629–634.

Morris, C.W., Boddy, L., and Allman, R. (1992) Identification of basidiomycete spores by neural network analysis of flow cytometry data. *Mycological Research* **96**, 697-701.

Morris, C.W., Boddy, L. and Wilkins, M.F. (1994) Approaches to applying neural networks to the identification of phytoplankton taxa from flow cytometry data. In: Dagli, C.H., Fernandez, B.R., Ghosh, J. and Kumara, R.T.S. (eds), *Intelligent Engineering Systems Through Artificial Neural Networks,* Vol. **4**. ASME Press, New York, pp. 619–627.

Rose, D.A. (1993) ROTTERS - An expert key for the identification of wood-rotting fungi in culture. *Binary* **5**, 9-12.

Schindler, J., Paryzek, P. and Farmer, J. (1994) Identification of bacteria by artificial neural networks. *Binary* **6,** 191-196.

Smits, J.R.M., Breedveld, L.W., Derksen, M.W.J., Kateman, G., Balfoort, H.W., Snoek, J. and Hofstraat, J.W. (1992) Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. *Analytica Chimica Acta* **258**, 11-15.

Weeks, P.J.D., Gauld, I.D., Gaston, K.J. and O'Neill, M.A. (1997) Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research* **87**, 203-211.

White, I.M. and Scott, P.R. (1994) Computer information resources for pest identification: a review. In: Hawksworth, D.L. (ed.), *The Identification and Characterisation of Pest Organisms*. CAB International, Wallingford, pp. 129-137.

Wilkins, M.F., Boddy, L. and Morris, C.W. (1994a) Kohonen maps and learning vector quantization neural networks for analysis of multivariate biological data. *Binary* **6**, 64-72.

Wilkins, M.F., Boddy, L., Morris, C.W. and Jonker, R. (1996). A comparison of some neural and non-neural methods for identification of phytoplankton from flow cytometry data. *CABIOS* **12**, 9-18.

Wilkins, M.F., Morris, C.W. and Boddy, L. (1994b) A comparison of radial basis function and backpropagation neural networks for identification of marine phytoplankton from multivariate flow cytometry data. *CABIOS* **10**, 285-294.

Woolley, J.B., and Stone, N.D. (1987) Application of artificial intelligence to systematics: SYSTEX - a prototype expert system for species identification. *Systematic Zoology* **36**, 248-267.

SESSION 2

KEY SYSTEMS

PLANT PARASITIC NEMATODES AND AUTOMATED TAXONOMIC TOOLS

By W.M. Hominick
CABI Bioscience, Bakeham Lane, Egham, Surrey TW20 9TY

1. Introduction

The intention of this part of the workshop was to explore which new approaches and technologies offer the best hope for automating systems or providing support for the identification of the more numerous and difficult invertebrates and microorganisms.

The view at CABI Bioscience is that plant nematodes lend themselves very well to computer-aided taxonomy because :

- While there are many new species to be discovered, and many species have been described, there are comparatively few that are economically important.
- The group is morphologically conservative (no wings; no legs; no sounds; no colour; minute size). Hence many drawings and photographs are necessary and these can be accommodated in the electronic format.
- Quantitative and qualitative characters are required and these can be provided in data sheets linked to descriptions.
- Dichotomous keys are currently the basic tool for identifying virtually all taxonomic groups of nematodes, and these lend themselves easily to computerised keys (see White and the use of TAXAKEY in this volume).
- Many groups are in a state of flux and keys do not exist. For these in particular, an excellent library and reference collection are required so that workers can sort out and assess particular specimens on site. The bibliographic material can be provided on a CD-ROM.

Our approach to identifying nematodes at IIP has always been pragmatic. Years of experience have shown that scientists, quarantine and advisory officers and other personnel concerned with nematodes place great faith in keys because they want a name at the end of the day. Hence, they will frequently see characters they want to see or need to see and if a couplet doesn't work, they will go back to make it work. Eventually, they arrive at a name and stop there. The number of incorrect identifications is of great concern to us. Our desire is to train people to realise that the work really only starts when a name is reached. The specimen needs to be checked against descriptions, and for that you need a library. One great advantage to an electronic key is that it can be designed so that the user is led immediately to a full text copy of the paper that describes the group or species as well as to linked databases providing additional information on the taxon.

In addition to keys, which may or may not exist, taxonomists, parataxonomists and other scientists require a library and copies of key publications to help in decision making. This is where there is a huge need in the BioNET-INTERNATIONAL LOOPs and a large opportunity for the electronic medium to play a key role in developing taxonomic resources, not only for nematodes, but for other groups as well. Such electronic bibliographic products, which can be tailored to specific requirements, can provide a foundation for developing taxonomic expertise.

This presentation discusses the present status of automated taxonomic tools for identification of plant parasitic nematodes, a group that is economically important, especially in the tropics, but with few trained taxonomists to deal with them. Perhaps for this reason, a number of electronic products are available or under development for release in the near future. Similar products could be developed for any group of organisms.

2. Root Knot Nematode Taxonomic Database

Root-knot nematodes, *Meloidogyne* species, are the most economically important and wide-spread plant parasitic nematodes in the tropics. They are also a taxonomically difficult group. The swollen females, which are the stages usually encountered on roots, display extreme morphological conservatism and some species show little host specificity. In the absence of a suitable key, Jon Eisenback, working at Virginia Polytechnic Institute and State University and an authority on root-knot nematodes, applied his skills with computers to create an up-to-date reference and teaching resource for nematologists and others interested in root-knot nematodes as a CD-ROM which has now been published (Eisenback, 1997).

The database is made up of a collection of over 450 articles, monographs and book chapters, 6 video clips and over 100 pictures, providing a unique reference resource on all aspects of root-knot nematode taxonomy.

Section headings include

- Morphology
- Diagnostic Keys
- References
- Classic Papers
- Techniques
- Distribution Maps
- Descriptions of Species
- Molecular Biology

Every document has been scanned to produce a PDF (Portable Document Format) file, which can then be read on-screen by the Adobe® Acrobat® Reader® software supplied with the disc. The video clips are run using QuickTime movie player, also supplied on the disc. Articles can be printed out for future reference or stored on the user's hard disk.

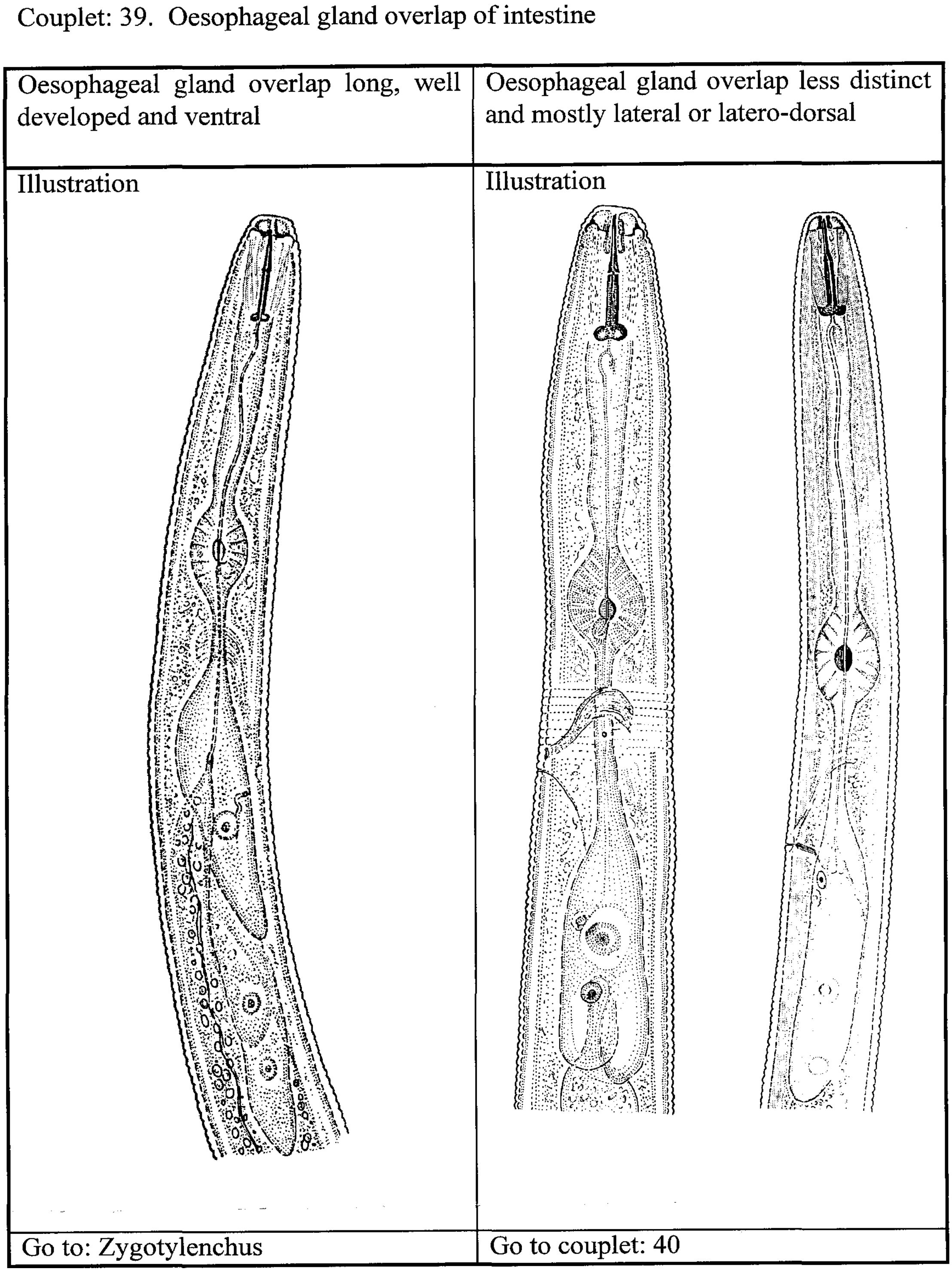
The unique feature of this CD-ROM is that for the price of a reference book, a small laboratory or under-resourced library can have available all the reference material required to work with the group, as well as added value information on geographical distribution, translations of classical papers, species lists, high quality micrographs and photographs, and short movies on techniques. There can be no question that products such as this provide a fundamental resource for a practising taxonomist or trainer.

3. Other Electronic Resources for Plant Nematodes

Apart from the root-knot nematode taxonomic database, there are a number of other resources either available or under development for release soon.

- ELECTRONIC KEY TO GENERA OF PLANT PARASITIC NEMATODES IN SOUTHERN AFRICA. This is a joint project led by D.J. Hunt of IIP, with colleagues of the Biosystematics Division of the Plant Protection Research Institute, Agricultural Research Council, Pretoria and utilising the expertise of CABI's Information Institute and Publishing Division. Commissioned by SAFRINET to help identify 55 genera of plant parasitic nematodes in SAFRINET member countries, it is intended for applied nematologists, extension workers, agricultural advisors, foresters, biologists and teachers at tertiary institutes. Users are expected to have sufficient experience to extract, mount and recognise plant parasitic nematodes. The dichotomous key (see Figure 1 for an illustration of a couplet) will be available in early 1998 and is based on the book by Kleynhans *et al.* (1996), with significant additional information from other sources, including unpublished information. In addition to taxonomic details, some information on hosts and geographical distribution will also be included.

Figure 1. Example of a couplet from the SAFRINET key to plant parasitic nematodes as seen on the VDU. Arrows and text will be added to emphasise the features mentioned.



Couplet: 39. Oesophageal gland overlap of intestine

Oesophageal gland overlap long, well developed and ventral	Oesophageal gland overlap less distinct and mostly lateral or latero-dorsal
Illustration	Illustration
Go to: Zygotylenchus	Go to couplet: 40

- CABI'S CROP PROTECTION COMPENDIUM. Module 1 is now available and focuses on crops and pests in SE Asia and the Pacific. This CD-ROM covers 1000 major pests and their natural enemies and includes quarantine pests. It features 150 crops of the world in 150 countries. While nematodes are only a small proportion of the pests covered, this module provides extensive information on 29 genera of plant parasitic nematodes. The Global version of the compendium is now in production and will be available in 1998, containing a larger section on plant nematodes.

- ILLUSTRATED GLOSSARY OF NEMATODE TERMS. This CD-ROM is being produced by J. Eisenback for publication by CABI in early 1998.

- CYST NEMATODE BIBLIOGRAPHIC RESOURCE. This will complement the root-knot nematode CD-ROM and together they will provide the basic reference and teaching resource for the two most economically important groups of plant parasitic nematodes in tropical and temperate regions.

- NEMABASE. This database has recently been launched on floppy discs and on the Internet. Funded by public money and developed by Ferris *et al*. (1997), it is a database on nematode/plant interactions. It contains information on 6100 plant taxa (including higher taxonomic information, geographic origin, growth habit and use), 3900 major plant-parasitic nematodes to the race level (including details of higher taxonomic information) and 38000 interactions (with details of the nature of each plant and nematode interaction, the constraints of the records and the source and quality of the data). The information has been extracted from nearly 5000 articles in 6 core journals from the turn of the century. Approximately 70% of the available data on plant and nematode interactions is contained in the database. Unfortunately the project ran out of funds and the database can easily crash, but this will be rectified in subsequent releases. It can be scanned on the Internet at www.ipm.ucdavis.edu/NEMABASE/ .

- INTERNET. A large amount of information is available by searching the Internet, but as for much information on the net, its accuracy and usefulness (much is basic data with little add-on value), needs to be questioned. An excellent start for high quality information is from the Nematology Department at UC Davis (www.ipm.ucdavis.edu).

4. The Future

There is a view that a dichotomous key is not worth converting to an electronic format. A simple dichotomous key that merely duplicates a "paper" key holds no advantage and may actually be less user-friendly than a paper version. However, using current technology it is possible to design keys, based on a dichotomous structure but incorporating data sheets linked to each taxon, such as the SAFRINET key described above. These can provide information on host plants, biology, distribution, links to original publications, glossaries, large numbers of colour photos and line drawings, databases and maps. Such keys are much more powerful than paper versions can ever be.

As the preceding sections show, the technology necessary to produce electronic taxonomic tools exists now. Moreover, it is not restricted to nematodes, but can be applied to any taxon. The main costs are for staff time, assuming the computers, scanners and CD-ROM production facilities are available. Undoubtedly the principal cost is for the research necessary to produce data sheets, gather references, produce photographs and drawings, interpret the literature and produce a key. The other major cost is for time to convert the research into an electronic, interactive format. For example, the SAFRINET project will take at least 8 person-months. However, it is based on a book which was funded by public funds through the PPRI and represents years of manpower and accumulated expertise. The important point to note is that much of the research required has been done and can be converted into a value-added format for computer-aided taxonomy. The limiting factor is resources to put taxonomists together with computer-literate people to generate the products.

There is another important point to bear in mind before we get carried away by our enthusiasm. That is, while we have the expertise and technology, and we know that these tools are needed, we can't do everything. We need to be selective. Just because something can be done doesn't mean that it should be done. Priorities need to be set, based on what is needed, who will use the information, and where it will be used. This is probably a more difficult challenge than actually producing the electronic tools. Given that funds are always restricted, how can they be applied for maximum effect? There is no question that BioNET-INTERNATIONAL needs automated taxonomy to achieve its objectives. Now that it is apparent that the technology to facilitate this is available now, the global BioNET-INTERNATIONAL network of biosystematists needs first to assess requirements and then set priorities. Implementing them will be comparatively straightforward.

5. References

Eisenback, J.D. (ed.) (1997) *Root-knot Nematode Taxonomic Database.* CAB International:, Wallingford. CD-ROM.

Ferris, H., Caswell-Chen, E.P. and Westerdahl, B.B. (1997) *NEMABASE - A database of the host status of plants to plant-parasitic nematodes.* Department of Nematology, University of California, Davis.

Kleynhans, K.P.N., Van den Berg, E., Swart, A., Marais, M. and Buckley, N.H. (1996) *Plant Nematodes in South Africa*. Plant Protection Research Institute, Pretoria.

THE USE OF AUTOMATED TAXONOMIC TOOLS IN AN APPLIED CONTEXT:
THE CROP PROTECTION COMPENDIUM

By P.R. Scott
CAB International, Wallingford OX10 8DE

1. Introduction

This paper takes a single representative example of a novel electronic tool, the Crop Protection Compendium, to illustrate how automated taxonomic systems can take their place within a multimedia compilation with a rigorously applied purpose in pest management. While the Compendium idea was developed as 'a new concept in meeting information needs for the developing world', it also has much to offer to users in developed countries and specifically to those concerned with identification of pests, diseases and weeds.

The Compendium illustrates the benefits than can accrue from extensive international cooperation in a major project, and also the principle of shared ownership through a Development Consortium. The paper also describes how these came about.

2. Where the Idea Came From

In 1989 CABI, CTA and FAO co-organized an International Crop Protection Information Workshop (Harris and Scott, 1989), which was attended by delegates from 32 countries in the developed and developing world. Delegates expressed profound concern that delivery of information to support effective crop protection in developing countries was quite inadequate, and in strong contrast to its availability in the developed world. There was also a sense of expectation that the incipient revolution in information technology should play a strong role in lessening this inequality. CABI had prepared for the Workshop a demonstration CD-ROM containing abstracts about a number of crop pest and pathogen species, text and images from published data sheets about them, and images of published distribution maps. These were connected to a list of names and synonyms of the species, allowing the user to 'hot-link' between related information on the disc. Concepts such as this excited the delegates and caused them to recommend concerted action to compile information on pests and diseases in a standardized way, and to exploit the emerging capabilities of information technology to deliver it, especially to developing countries.

This concept of standardized pest information, delivered in electronic form, was elaborated through the use of relational database technology. The European and Mediterranean Plant Protection Organization (EPPO) had pioneered the concept of relational databases in which pest and pathogen species, crops and countries were the categories to be related (Smith and Smith, 1991). This relational structure had been taken up by the FAO Regional Office in Trinidad in the development of a practical computerized tool for recording what pest species were present where, and on what crops, with notes on biology, economic impact and control (Schotman, 1989).

The relationality of the data provided an important vehicle for a diagnostic dimension to the concept that was emerging. CABI developed the concept further, under the name of the Electronic Compendium for Crop Protection, and received support for a more detailed study and evaluation from the Australian Centre for International Agricultural Research (ACIAR).

3. Feasibility Study

ACIAR funded a Feasibility Study on the Electronic Compendium concept. The Study opened in 1992 with a small international workshop of 14 people who considered the needs of a range of users and developed an outline structural and navigational scheme for a working Compendium. A Consultant (Charles Schotman) prepared a more detailed specification, assessed resource needs, and developed a prototype demonstration that could be shown to potential users. The demonstration proved to be a powerful tool in engaging attention, and was extensively used in the conduct of a survey of user needs, focusing on South East Asia and the Pacific. This User Needs Survey produced an enthusiastic response. Table 1 illustrates the wide range of types of professional user identified, with an indication of their relative interest in using the Compendium if it were available.

Table 1. Assessment of the suitability of the Crop Protection Compendium to potential user groups, based on a survey of 126 institutions in the Asia-Pacific Region.

USER GROUPS	**(a)**	**(b)**
Research scientists	93	91
Pest managers	100	88
Quarantine officers	100	79
Lecturers	81	78
Extension officers	100	77
Regional organizations	80	71
Research managers	64	69
Plant breeders	67	62
Undergraduates	-	63
Agrochemical industry	78	63
Industrial crop companies	-	52
Farmers and growers	-	51
Non-governmental organizations	-	51
Policy makers	58	49
Donors and development assistance agencies	54	49
Food industry	-	43

(a) self assessment in percentage of maximum possible score
(b) overall assessment in percentage of maximum possible score

In addition to a prototype demonstration and a User Needs Survey, the Feasibility Study led to the development of an outline Specification and a Project Proposal to develop a working Compendium. The Proposal envisaged the compilation of a Compendium with global coverage in a stepwise fashion, the first step being to develop a Module focusing on approximately 1000 pests ('pests' includes pathogens and weeds) that are of particular importance in South East Asia, southern China and the Pacific.

4. Development Project

The two-year Development Project to compile this first Module of a global Crop Protection Compendium occupied the years 1995 and 1996. It was budgeted at US$1.5M. The core task was to prepare standardized, illustrated data sheets on 1000 pest species, on 150 crops and on 150 countries. The strategy was to engage a specialist for each species or group of species and to outsource the work to a total of several hundred such specialists. A Project Coordinator and a team of editors and other staff were located at CABI Headquarters in Britain, supported by a Regional Coordinator at CABI's Regional Office in Malaysia. Data sheets were commissioned, edited to agreed standards by the editorial team, and then verified for consistency and balance by a small group of specialists.

Meanwhile software was written by a Consultant and by CABI staff, based on the software prepared for the prototype module. Its foundation was a relational database system, reflecting relationships between pest species, crop species, countries, symptoms and plant parts. A friendly user interface was developed to allow navigation through the system on the basis of these relationships, supplemented by dynamic hyperlinks. The software included a Geographic Information System (GIS) for generating maps of the distribution of pests and crops, a hierarchical taxonomic structure, a bibliographic database and retrieval system, a hyperlinked glossary, a module for presenting statistical information about crops and commodities.

An important part of the software development was the inclusion of a series of interactive diagnostic keys that had been developed separately (White and Sandlant, 1998). The open architecture of the Compendium allows free-standing systems such as CABIKEY and TAXAKEY to be linked to it without difficulty. The relational structure of the data allows a simple approach to identification through narrowing down by country, crop, symptom, growth stage, etc. This relationality is a further important dimension of the Compendium that gives it some of its diagnostic power.

5. Development Consortium

The project was resourced by a Development Consortium of 20 organizations (Table 2), each of which committed funding in units of $50,000 over the two-year period of the project. The principle of shared ownership of the project and commitment to its completion proved valuable. It successfully embraced both public-sector and private-sector organizations in a common endeavour, and allowed in-kind commitment of resources as well as financial support.

Consortium Members benefit through public identification with this innovative project, influence on the project and its output, the gearing effect of cooperative funding, mutual confidence that the project would deliver an output to schedule, and privileged access to the finished product. The fact that CABI had engaged 20 Consortium Members who jointly provided full development funding for the project provides a measure of demand for the Compendium. This evidence is strengthened by the advance purchases of copies of the product made by many Consortium Members.

6. Sustainability

CABI has published the Compendium on CD-ROM (CABI, 1997), using a dual-pricing formula that allows customers in developing countries a 75% discount on the market price of the product in the developed world.

CABI has also undertaken to publish an updated edition at least annually. Revenue from sales will be recycled, so that maintenance of the product is assured. This provides a mechanism for sustainability of the Compendium, which is an important consideration for its stakeholders including Consortium Members and purchasers of the product. CABI's status as an international organization with a mission to support the needs of developing countries, and as a not-for-profit electronic publisher, gives it a favourable position to complete this model of:

- Development costs covered by a Development Consortium.
- Sustainability then ensured through the market.

Table 2. Founder Members of the Development Consortium for CABI's Crop Protection Compendium.

CAB International
AgrEvo, Germany
Asian Development Bank (ADB)
Australian Centre for International Agricultural Research (ACIAR)
Canadian International Development Agency (CIDA)
Cyanamid, USA
Danish Government Institute of Seed Pathology/Danish International Development Agency (DGISP/DANIDA)
Department for International Development (DFID), UK
DowElanco, USA
DuPont, USA
Gesellschaft für Technische Zusammenarbeit (GTZ), Germany
International Development Research Centre (IDRC), Canada
International Rice Research Institute (IRRI), Philippines
Monsanto, USA
Novartis Crop Protection, Switzerland
Pioneer Hi-Bred, USA
Rohm & Haas, USA
Sumitomo Chemical Company Limited, Japan
Swiss Agency for Development and Cooperation (SDC)
United Nations Development Program (UNDP)
United States Department of Agriculture: Animal & Plant Health Inspection Service (USDA-APHIS)
Zeneca Agrochemicals, UK

7. Structure and Content of the Crop Protection Compendium

Delivery of this paper included a live demonstration of the Crop Protection Compendium. The requirements are an IBM-compatible PC (minimum 486DX) with Windows 3.1 or Windows 95, a minimum of 8Mb RAM (16Mb recommended), 20Mb available on the hard disc, a CD-ROM drive, and a VGA display (preferably supporting 256 colours). The demonstration was used to illustrate the following principal components of the Compendium system and its user interface.

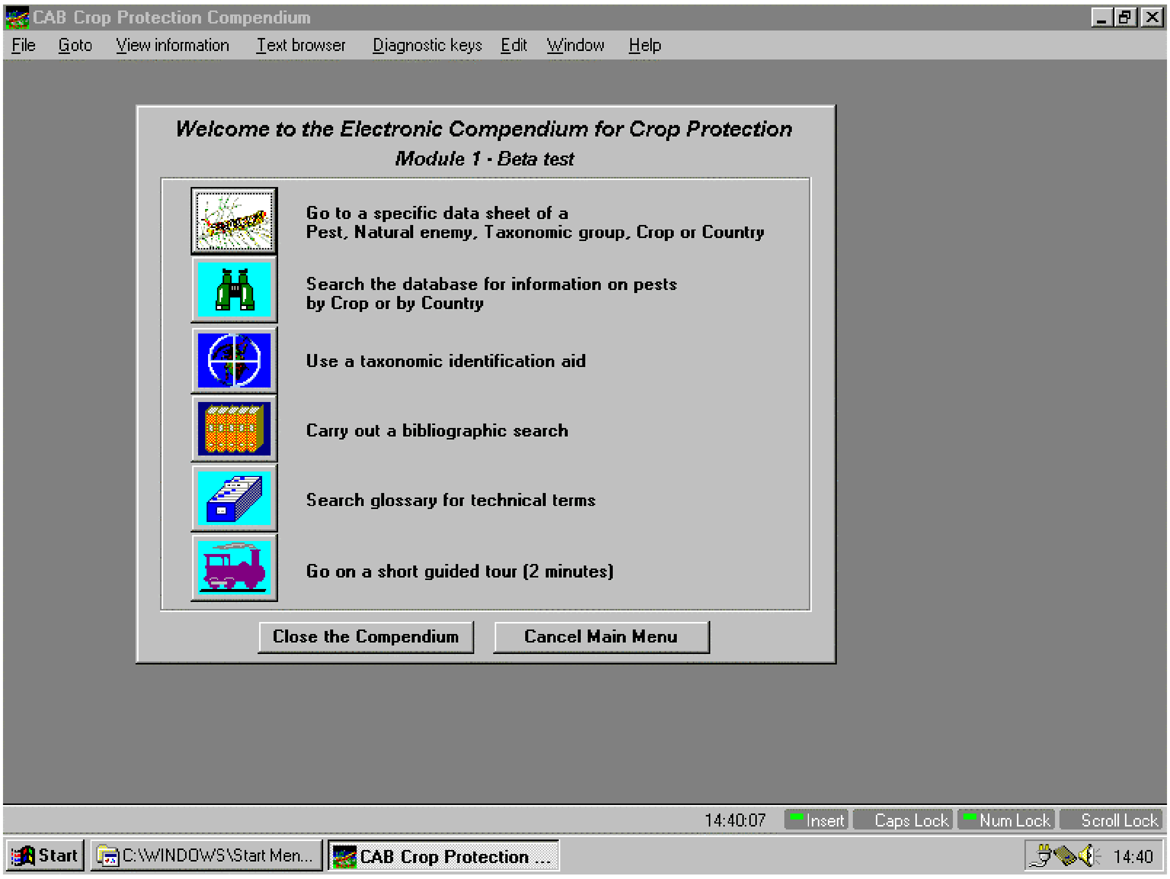
7.1 Opening menu

Figure 1 shows the self-explanatory menu presented to the user when the Compendium is launched.

7.2 Data sheets on pests and natural enemies

The first menu button gives access to a long list of names of pests and natural enemies, including scientific names, synonyms, and common names in a number of languages. The list can be scrolled and a name selected, or an incremental search facility allows the target name to be reached by keying in its opening letters. In this first Compendium Module, data sheets are provided for approximately 1000 species of pests and their natural enemies that are of particular importance in South East Asia, southern China and the Pacific. Each was prepared by one of more than 400 specialists (more than 25% from the Asian region), and then edited and validated by a core team of editors and crop protection professionals. The viewpoint is always global, since the components of this first Module will also be present in the full Crop Protection Compendium which will have global coverage.

Figure 1. Crop Protection Compendium: opening menu.



Welcome to the Electronic Compendium for Crop Protection

Module 1 - Beta test

Go to a specific data sheet of a
Pest, Natural enemy, Taxonomic group, Crop or Country

Search the database for information on pests
by Crop or by Country

Use a taxonomic identification aid

Carry out a bibliographic search

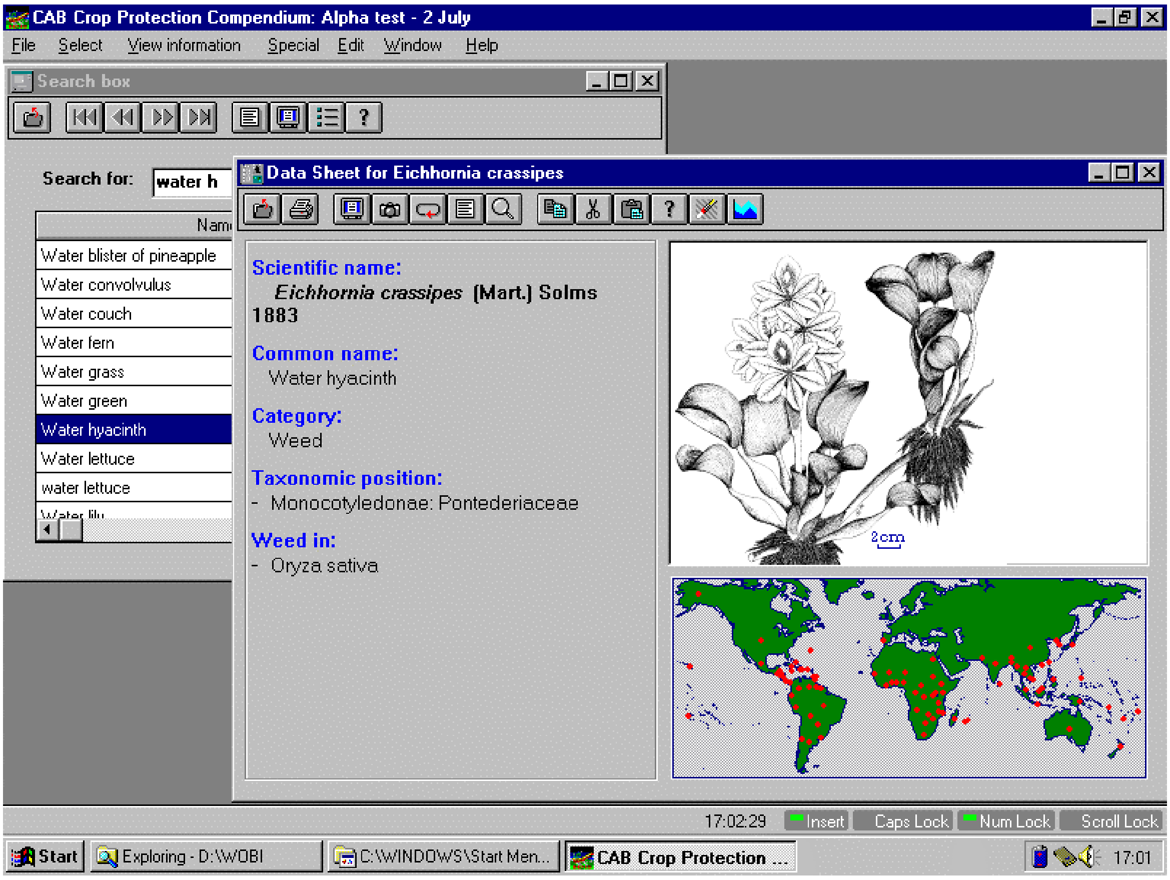
Search glossary for technical terms

Go on a short guided tour (2 minutes)

The opening page of the selected data sheet (Figure 2) provides a brief textual description of the pest, a picture, and a small distribution map. Menus or buttons then give access to the details. The core of the data sheet is a series of sections, individually accessible, covering:

Names and taxonomic position
Host range / habitat
Geographic distribution
Biology and ecology
Seedborne aspects
Natural enemies
Economic impact
Phytosanitary significance
Symptoms
Morphology
Similarities with other pests
Detection and inspection
Diagnosis
Control
References.

Figure 2. Crop Protection Compendium: example of a pest data sheet



Scientific name:
Eichhornia crassipes (Mart.) Solms 1883

Common name:
Water hyacinth

Category:
Weed

Taxonomic position:
- Monocotyledonae: Pontederiaceae

Weed in:
- Oryza sativa

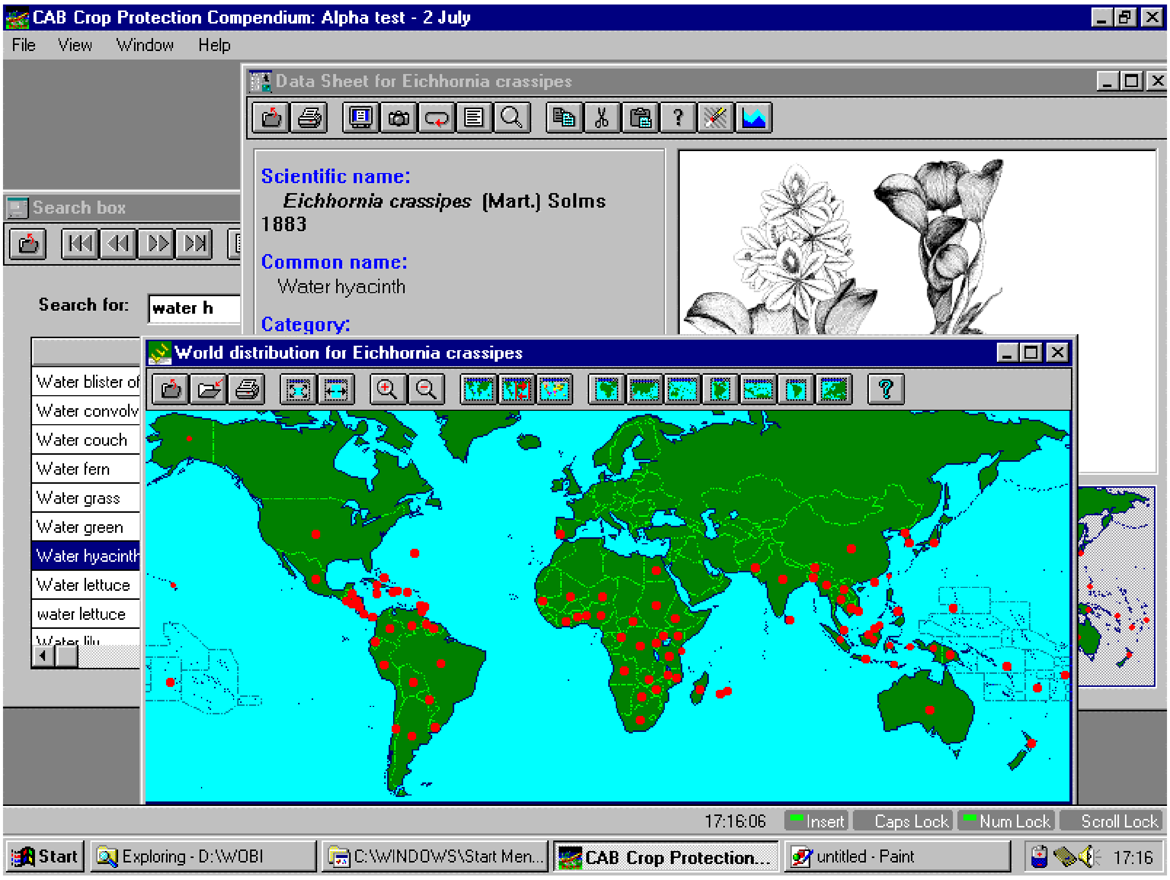
Some of these are further subdivided into sections accessed by a further menu. The section on Control, for example, covers the several components of Integrated Pest Management (IPM) including:

Biological control
Chemical control
Cultural and physical control
Host-plant resistance
Regulatory control.

Each pest data sheet also provides a facility for users to keep their own notes, personally (restricted to their own PCs) and corporately (accessible to other users over a network). All the text, whether in a prepared data sheet or added by the user, can be searched, copied and downloaded. Furthermore any word or string of words can be used to provide a 'soft link' to another data sheet. Thus in the data sheet on water hyacinth (*Eichhornia crassipes*), the words *Monochoria vaginalis* occur in the text; they can be selected and used as a soft link to load immediately the data sheet on this similar species, pickerel weed.

A button gives access to a full-screen global distribution map, and further buttons zoom in to continents or smaller areas (Figure 3). These maps are generated dynamically from the relational database of pest species and countries. Individual points on a map can be clicked with a mouse to show the detailed records. Maps can be overlaid, for example to show the distribution of a pest and its host crop together.

Figure 3. Crop Protection Compendium: example of a pest distribution map.



Scientific name:
Eichhornia crassipes (Mart.) Solms 1883

Common name:
Water hyacinth

Category:

Search for: water h

- Water blister of
- Water convolv
- Water couch
- Water fern
- Water grass
- Water green
- Water hyacinth
- Water lettuce
- water lettuce
- Water lily

Pictures are accessed through buttons and menus, via lists of captions and previews of reduced images (Figure 4).

Other items accessible from menus include taxonomic trees and lists of natural enemies.

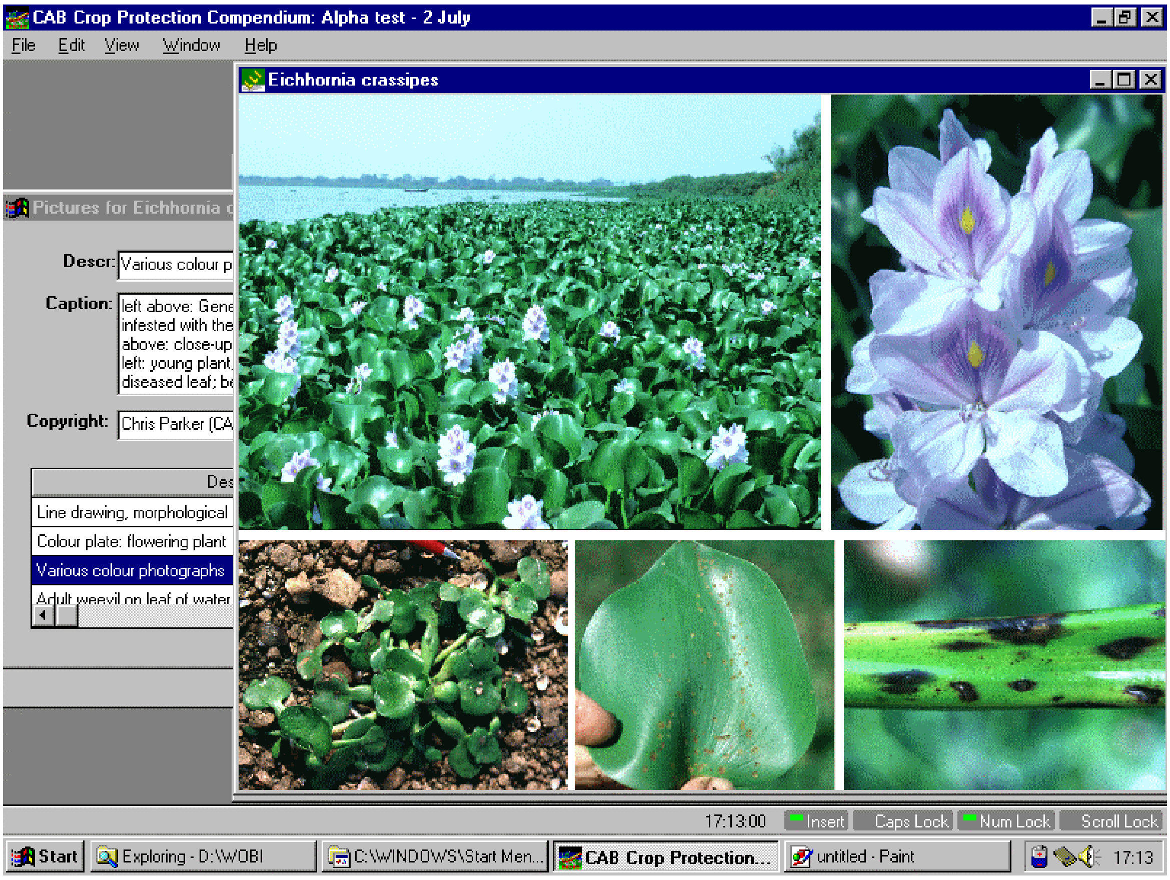
7.3 Data sheets on crops and countries

The opening menu also gives access to data sheets on 150 crops and 150 countries. The structure of each of these is also based on an opening page with summary text, picture and map. Again, menus or buttons give access to the detail.

Crop data sheets typically contain components on:

Names and taxonomic position
Habitat
Geographic distribution
Biology and ecology
Botanical description
Uses
Production and international trade
Agronomic aspects
Pests and diseases
Genetic resources / breeding
References.

Figure 4. Crop Protection Compendium: example of illustrations.



CAB Crop Protection Compendium: Alpha test - 2 July
File
Edit
View
Window
Help
Eichhornia crassipes
Pictures for Eichhornia c

Descr: Various colour p

Caption: left above: Gene infested with the above: close-up left: young plant, diseased leaf; be

Copyright: Chris Parker (CA

Des
Line drawing, morphological
Colour plate: flowering plant
Various colour photographs
Adult weevil on leaf of water

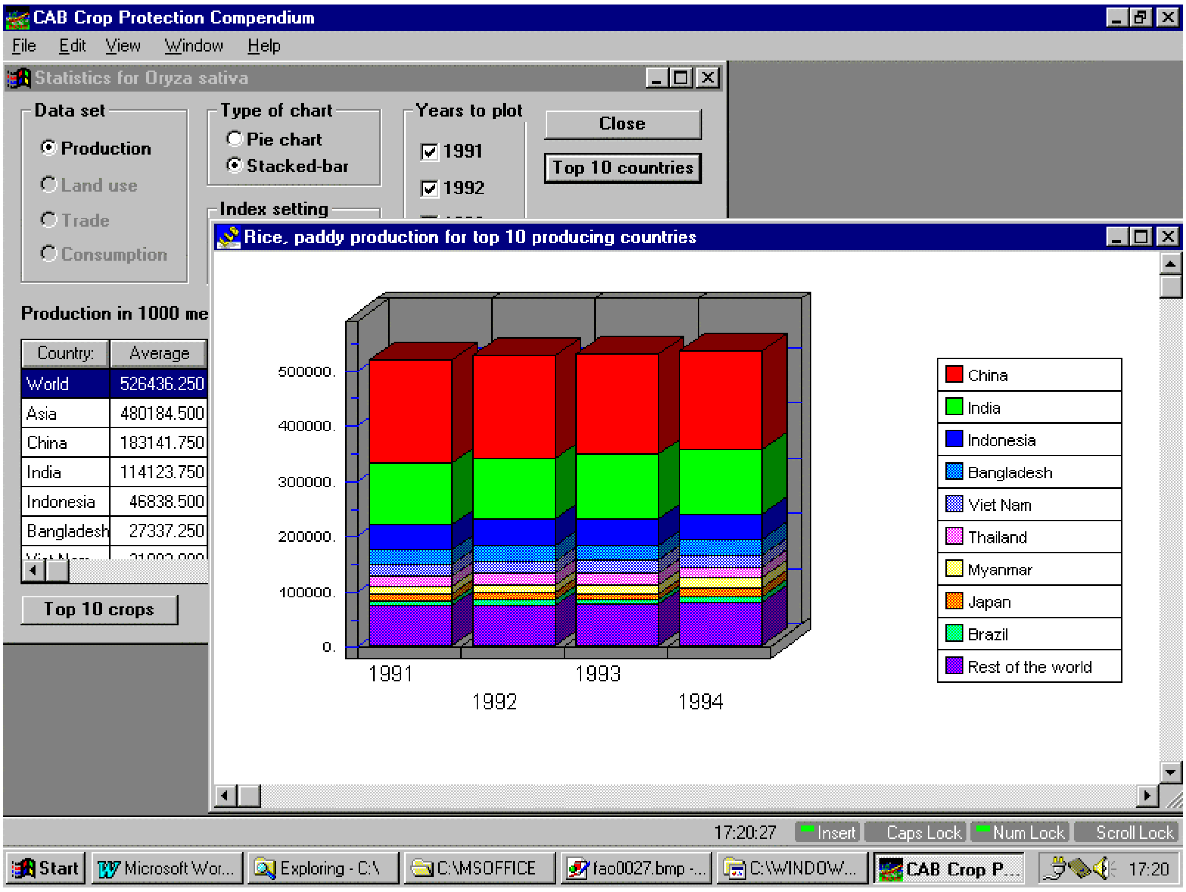
Much of the information is derived from existing resources for which CABI has negotiated access, notably the Plant Resources of South East Asia (PROSEA) initiative (Lemmens, 1989).

Country data sheets include brief information about climate, land area, population, etc., and details of national and regional plant protection organizations. Pesticide usage data are available for selected Asian countries, from an authoritative source, broken down by crop and type of pest.

All text can be handled and used for soft linking as for the pest data sheets, and facilities are again available for users to make their own notes on an individual crop or country at the personal or corporate level.

Buttons give access to distribution maps and pictures, with similar facilities to those for pests. An extra button is used to present statistical information, drawn from FAO and other sources, in tabular or graphical form (Figure 5).

Figure 5. Crop Protection Compendium: example of statistical information.



Statistics for Oryza sativa

Data set
● Production
○ Land use
○ Trade
○ Consumption

Type of chart
○ Pie chart
● Stacked-bar

Years to plot
☑ 1991
☑ 1992

Production in 1000 me

Country:	Average
World	526436.250
Asia	480184.500
China	183141.750
India	114123.750
Indonesia	46838.500
Bangladesh	27337.250

7.4 Identification: searching for pest information by crop or country

The second button on the opening menu opens a dialogue that allows pest species to be listed based on their match with the criteria by which the underlying relational database is classified:

Pest group
Crop
Plant parts
Symptoms
Stage affected
Country.

This is a simple but powerful method of approaching pest diagnosis by rapidly narrowing down from the full list of several tens of thousands of names to the handful that match a set of criteria. The resulting short list provides direct access to full data sheets on each of the species listed.

7.5 Diagnostic keys

The opening menu provides access to a set of diagnostic keys. Eventually, the aim is to allow diagnosis to within a few species for any pest in the Compendium. In the first Module there are keys for:

Arthropod orders
Heteroptera families
Homoptera families
Lepidoptera families (adults)

Lepidoptera families (caterpillars)
Fruit flies: Asian species of *Dacus* and *Bactrocera*
Beetle families
Weed species currently included in the Compendium
Nematode species currently included in the Compendium.

These computer-aided identification tools include CABIKEY multi-entry keys and TAXAKEY dichotomous keys, all with illustrations (White, this volume).

A result derived from an identification session is automatically fed back to the Compendium, which then loads the relevant data sheet. If this is for a family, for example, the Compendium can compile a list of all species in that family that are relevant to crop protection. Further progress can then be made by invoking the narrow-down procedure, described in the previous section, to identify just those species that are present in a specified country, on a specified crop, etc.

7.6 References

A large bibliographic database and search tool are included in the Compendium. The database contains all the references listed within individual data sheets, plus many others on IPM compiled from international, regional and national sources and including non-conventional or 'grey' literature. In total there are some 60,000 references, most with abstracts. Those that are cited in data sheets are internally connected with their citations. All can be accessed with the built-in retrieval system which provides the normal Boolean search facilities and handling of search histories, plus a simple natural-language search tool.

7.7 Glossary

A glossary module of several thousand terms used in pest management can be accessed from the opening menu, or by soft-linking from text within any part of the Compendium. The contents have been compiled from several sources, on the understanding that their owners will receive feedback from the experience of users of the Compendium. The glossary module includes a search system and internal cross-referencing. The system is flexible and can accommodate any kind of information for which simple alphabetic indexing is appropriate. This includes pesticide information, for example, and data are listed for all active ingredients from the *Pesticide Manual* (Tomlin, 1994), including:

Mode of action
Uses
Phytotoxicity
Formulations
Compatibility
Trade names
Mixtures
Mammalian toxicology
Ecotoxicology
Environmental fate.

7.8 Guided tour

The opening menu also offers a guided tour of the main components of the Compendium, to help familiarize a new user with its capabilities.

8. The Next Steps

CABI's team that developed Module 1 of the Crop Protection Compendium is engaged in preparing a Global Module with substantially extended coverage, to approximately 1800 species of pests, diseases, weeds and their natural enemies, ready for publication in 1999. This initiative of globalization has been taken by the Development Consortium, whose membership has been extended so that it can cover the full development costs. Including Module 1 these amount to US$3M over 4 years.

The relational structure of the Compendium means that an extended version inherently retains its capacity to provide a diagnostic capability through the narrow-down approach. In addition, a pathogen diagnostic aid is being added, through extension of the range of symptoms that are relationally indexed.

The user's Web Browser can be launched from within the Compendium, giving access to a Crop Protection Compendium Home Page. This can be used to provide links to other relevant sites, and to present recent information and news. This represents the first step towards delivery of the content over the Web, which is a likely option for the future.

9. The Electronic Compendium as a Knowledge Platform that Places Diagnostic Aids in an Applied Context

This paper has focused on a single knowledge platform as an example of a tool that can place diagnostic aids in a firmly applied context. It illustrates the power of information technology to provide that context, and is a tangible sequel to the sense of expectation at the 1989 International Crop Protection Information Workshop.

Developments of this sort provide a test of the value of information technology in supporting applied tasks. The signs are that the Crop Protection Compendium is meeting this need, since it has received many recommendations, and its diagnostic component is often mentioned. Its original focus on the needs of developing countries still presents some challenges in ensuring that the tool is affordable and that the required infrastructure is in place. This challenge has been grasped by the Development Consortium that supports the Crop Protection Compendium project. Several of its Members have pre-purchased copies of the product with a view to their installation in developing countries. One of these, the Asian Development Bank, insisted on complementing its contribution to the development project ($200,000) with a substantial further contribution ($100,000) dedicated to providing hardware and training in six Asian countries. A feature of this has been the engagement of trainees in the future of the Compendium, by encouraging maximum feedback to help improve future editions. The user notepad facility within the Compendium, which provides one mechanism to facilitate this, is a tangible signal to users that they can play a valuable role, as stakeholders in the project, in ensuring its continuation and improvement.

10. References

CABI (1997) *Crop Protection Compendium: Module 1*. Multimedia CD-ROM. CAB International, Wallingford.

Harris, K.M. and Scott, P.R. (1989) *Crop Protection Information: an International Perspective*. CAB International, Wallingford.

Schotman, C.Y.L. (1989) New pest records. *Circular Letter, FAO Regional Office for Latin America and the Caribbean*. FAO: Santiago.

Lemmens, R.H.M.J., Jansen, P.C.M., Seimonsma, J.S. and Stavast, F.M. (1989) *Basic List of Species and Commodity Grouping.* Version 1. Wageningen, Netherlands: PROSEA Project.

Smith, I.M. and Smith, B.C. (1991) PQ - the EPPO data base on plant quarantine pests. *EPPO Bulletin* **21**, 211-218.

Tomlin, C. (1994) *The Pesticide Manual*. British Crop Protection Council, Farnham.

White, I.M. and Sandlant, G. (1998) Computerized insect identification: a comparison of differing approaches and problems. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 261-272.

CABIKEY AND TAXAKEY

By I. White
Formerly of International Institute of Entomology, 55 Queen’s Gate, London SW7 5DR

ABSTRACT OF PAPER

The advantages and disadvantages of computerised dichotomous and multiple-entry keys are discussed. While both approaches allow pictorial presentation and space for explanation of terms, the multiple-entry key also has better retention of information, greater user efficiency, and the power to allow users some control of the path to an identification. A theoretical comparison was made of how well each approach represents an information space (a data matrix of taxa by characters); multiple-entry keys can represent the entire information space; conversely, a dichotomous key could in some cases contain under 10% of that information. For example, a dichotomous key to mosquito genera was found to contain only 17% of the possible information that the author would have had to consider in some way during the construction of that key. Multiple-entry keys also have greater user efficiency in terms of the mean number of questions that have to be answered to achieve an identification. An experiment comparing dichotomous and multiple-entry keys to 120 species of tephritid fruit flies was carried out. A selection of 20 pest species was run through each key and it was found that 179 decisions were required to use the multiple-entry key compared to 244 with the dichotomous key, an increase of 36%. In general the multiple-entry key approach is to be preferred, both for its greater information content and user efficiency. However, in some circumstances there may be good reason to simply convert an existing printed dichotomous key into computerised form in order to gain the advantages of improved presentation and links to other information. Furthermore, multiple-entry keys can be difficult to apply in some circumstances, in particular where the taxa are of genus level or above. For example, some mosquito genera appear to be unnatural groups of species and some questions have to contain combinations of two or more characters. A subjective comparison of the application of the multiple-entry key approach to species (fruit flies), genera (mosquitoes) and families (beetles) suggests that this technique gets increasingly more difficult to use with increasing taxonomic level.

Author’s note:The full paper has been published in full as: White, I.M. and Sandlant, G. (1998) Computerized insect identification: a comparison of differing approaches and problems. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 261-272.

SESSION 3

VR AND TRAINING

THE POTENTIAL OF QUICKTIME VR OBJECT MOVIES OF INVERTEBRATES AS TAXONOMIC TRAINING AIDS

By W. Parker
ADAS, Woodthorne, Wolverhampton, WV6 8TQ

C. Dent
Spacexploration Ltd, 51 Compayne Gardens, West Hampstead, London NW6 3DB

1. Introduction

BioNET-INTERNATIONAL has set itself a wide-ranging and challenging task which encompasses a number of key objectives. These essentially revolve around developing technologies for information transfer from the main world centres of taxonomic excellence, the re-habilitation and development of reference collections, and the development of new technologies to make biosystematics accessible throughout the developing world. Inherent in all these objectives is a strong emphasis on the need to train both specialist and non-specialist taxonomists in the developing world to a level where they start to make a significant contribution to the cataloguing and management of biodiversity in their respective countries.

Modern multi-media techniques are revolutionising the way in which information can be presented and disseminated. Provided such technology can be made available to developing countries, it ought to play a significant role in aiding the technology transfer and training objectives of BioNET-INTERNATIONAL. However, it is crucial that the applications which are developed are focused on clear, identified needs arising from within the BioNET-INTERNATIONAL LOOPs. This paper briefly introduces one aspect of modern multi-media technology, object movies created using QuickTimeVR, which could be tailored to fit taxonomic training needs on a number of different levels. These range from providing training in basic recognition of, for example, different orders of insects to integration into more sophisticated identification packages such as CABIKEY.

2. What are QuickTimeVR Object Movies?

An object movie is a very careful photo-montage of a real object created from a series of still photographs taken from all angles. The pictures are then stitched together using QuickTimeVR to create a single image with a difference - the finished image allows the object to be turned round or over, and provided the right pictures have been taken, it is possible to zoom in to a considerable level of detail. In addition, areas of the image can designated as 'hot-spots' allowing labels to appear as the mouse rolls over the area. Hyper-links can easily be created to other text, images or sound clips. Thus object movies are ideally suited for incorporation into other electronic training media, whether true multi-media or just embedded into other electronic documents. The software required to run the movies is readily available as it is used widely in many commercial CD-ROMs, or it can be downloaded from Apple\'92s website on the Internet.

3. Object Movies as Taxonomic Training Aids

A critical step to achieving the longer term aims of BioNET-INTERNATIONAL, is to have high quality, accessible training courses and training aids made available where they are most needed. It is precisely in this area that object movies perhaps have most to offer. Imagine yourself as a putative taxonomist struggling with inadequate equipment to teach yourself about invertebrate identification. You might have no microscope, or only a poor one with inadequate lighting. You might have a limited supply of specimens, some of which might be delicate or rare. You might not have any relevant text books, or they might be in a foreign language. Worst of all, you may not have any pictures to guide you on what you should be looking for, particularly with regard to what some of the taxonomic terminology is actually referring to! Everybody knows that a picture is worth a thousand words, and where some or all of the nightmare scenario outlined above is a reality, the use of object movies integrated into a proper education package could make a real contribution:

- Train on actual material rather than rely on idealised one-dimensional drawings or too-perfect rendered images.

- Training without microscopes or video equipment. Provided the photographs are taken in sufficient detail and magnification, very high levels of detail, certainly sufficient for most taxonomic training purposes can be obtained without the need for microscopes. It would be possible to learn the process of using a taxonomic key without the frustration of having to obtain and handle an actual specimen.

- Zoom in to look at key features such as tarsal segments or bristles without worrying about breaking them off or struggling to turn the specimen to the right angle under the microscope.

- Work on rare or delicate specimens from around the world, or build-up a customised reference collection of specimens which have been photographed and processed into object movies.

- Object-movie reference collections could either reside on CDs, or could be held centrally on a database on the Internet and accessed directly. A centralised system would have the advantage of being easily added to and updated, although it would of course require users to have Internet access.

- The language and terminology barriers can be overcome by ensuring the text associated with the object movie is in the appropriate language, or played as a sound clip for those with literacy difficulties.

Thus although object movies are not a solution in themselves, they have the potential to make a contribution to the technology transfer process, from basic training in recognition of important agricultural or phytosanitary pest species, through to more detailed training in taxonomic features and techniques.

4. Limitations

There are some technical limitations to the use of object movies. Primarily, a computer, preferably with a CD-ROM drive (although this is not essential), is required to run the necessary software. Access to computers may be severely limited in some parts of the world. This will be a common problem to all innovative IT technologies, the importance of which BioNET-

INTERNATIONAL will need to assess. The size of object that can be adequately photographed is also a limitation. While medium-sized Coleoptera such as cereal leaf beetle (*Oulema melanopus*), small beetles such as pollen beetles (*Meligethes* spp.) and larger Diptera (e.g. *Sarcophaga* spp.) have been successfully photographed, very small softer-bodied invertebrates such as some Homoptera and Acari would be much more technically challenging to photograph in sufficient detail. Insects with large wing-spans (particularly Odonata) might also be problematic in terms of gaining adequate depth of field to keep the whole object in focus. However, these are technical problems which can be overcome and therefore should not be seen as severely limiting.

5. Conclusion

Achieving the overall aims of BioNET-INTERNATIONAL will only be possible through the integration of many different approaches to developing and disseminating biosystematic skills and techniques. There is room here for many different technologies, of which Quicktime VR object movies are only one. However, object movies have the advantage of relying on already proven techniques, and therefore require no further investment to make them useable. This technology can make a serious contribution to the overall BioNET-INTERNATIONAL objectives.

COMPUTER-BASED IDENTIFICATION : THE UNASKED QUESTIONS

By M. Edwards
33 George Street, Berkhamsted, Herts, HP4 2EG

1. Introduction

There is a wide range of computer-based tools and techniques available to help with species identification. Some of these tools are well developed, for example, the multi-access keys CABIKEY (White, this volume), INTKEY (Dallwitz *et al*., 1998) and PANKEY, and the use of hypertext and multimedia tools. Other techniques are at a more developmental stage, these include the use of image analysis (O'Neill, this volume) and neural networks (Boddy, this volume). Some of these techniques use conventional taxonomic characters, while others use unconventional data derived from image analysis (O'Neill, this volume), acoustic data (Chesmore *et al*., 1998; Chesmore, this volume) and flow cytometry (Boddy, this volume). Hence there is now a range of techniques which can be used to develop species identification tools, however, there remain a number of fundamental issues which have yet to be addressed, and must be addressed, if significant progress is to be made in providing tools to support species identification. This paper addresses two issues:

A. In order to use the available tools effectively, and to develop appropriate new tools, it is necessary to understand the circumstances under which identification is done, and the needs of those doing identification. This section raises many questions about species identification, but is able to provide few answers.

B. Given that there is this range of identification tools, guidance is needed to enable those doing identification, and those supporting identification, to decide which tools are most appropriate for which species groups and under which circumstances. This section is able to give some guidelines, but these will need to be modified when we find the answers to questions outlined in the preceding section.

2. Understanding the Identification Process

This section states a number of questions which need to be answered if research into, and development of, new species identification tools is to be effective.

2.1 Who and how many people do identification?

Three different groups of people can immediately be identified as being involved in species identification: specialised taxonomists, trained biologists, and parataxonomists (Alberch, 1993) and untrained biologists, but there may be others. There are estimates of the numbers of specialised taxonomists (e.g. Gaston and May, 1992), but not of the other groups. These different groups clearly have different levels of expertise and will therefore require different identification tools to address their particular needs. For example, an identification tool for parataxonomists and untrained biologists should include a substantial training component. Unless we know the proportion of members in the different groups, there is a danger of not developing tools which address the needs of the largest group of workers in species identification.

2.2 Why do people do identification?

People do identification for many reasons, including taxonomic studies, biodiversity and habitat studies, agricultural studies (e.g. pest control, crop management and plant quarantine) and medical studies (e.g. parasitology).

Different biological studies have different objectives. In many cases species identification is only a small component of the overall study. If the objective of the study is pest control, on reaching an identification the user expects to be given advice on treatment. This suggests that it should be possible to integrate identification tools within other advisory systems, as is the case with the Crop Protection Compendium (Scott, this volume).

2.3 Where is identification done (1)?

Species identification is undertaken world-wide in tropical, subtropical, temperate and polar regions, in both terrestrial and marine environments. The main issue in relation to geographical location is the likelihood of finding new species. In temperate terrestrial studies new species are the exception rather than the rule, whereas in the tropics new species are to be expected. This should be reflected in the tools provided. For example, failure to identify a European insect is likely to be due to an error made by the user of a tool, whereas failure to identify a tropical insect is quite possibly because it is a new species. A tool developed for a species group or a region where the possibility of new species is high, should be able to recognise when an identification cannot be reached and should warn the user of the possibility of a new species.

2.4 Where is identification done (2)?

Where people do identification will have major implications for what type of tools, especially hardware, are appropriate. Developments in laptop and handheld computers mean that such equipment can now be used in the field, so it is now appropriate to develop tools for groups (e.g. angiosperms) which are traditionally identified in the field. Biologists working in makeshift labs may have to cope with restricted or unreliable power supplies and restricted availability of equipment, consequently they may be obliged to use more basic tools and techniques for identification. In well-equipped laboratories, facilities will exist to use more sophisticated techniques, e.g. gel electrophoresis and electron microscopy. If this is the case then the results of such tests can be included in the identification tools. However, an important characteristic of some identifications is that a result is needed now. For example, a quarantine officer needs to be able to make a decision quickly, and cannot wait for the outcome of lengthy tests.

2.5 What level of detail is required?

Different studies require identifications to be done to different levels of detail. For example, in some ecological studies identification to OTU (operational taxonomic unit) is adequate. For other ecological, habitat and biodiversity studies identification to order, family or genus will suffice. Biologists working on these types of studies want tools which are not too detailed, or at least give them the option of finishing an identification at the appropriate level of detail. However, in studies relating to biological control and parasitology studies, as well as taxonomic studies, the identification must be exact to specific or sub-specific level, and tools are needed which provide this level of detail.

2.6 What tools would people doing identification in the field find useful?

Many of the tools currently available to assist with identification do not appear to have been developed in conjunction with those using the tools. We need to talk to experienced practitioners, both those doing identifications in the laboratory and those working in the field, to find out what sort of tools would assist them in identification. For example, field botanists might find a handheld computer, which combined the functions of a field notebook and a flora, useful. Such a tool would have to take into account the limited storage capacity and screen area of such computers.

2.7 How do people do identification?

If tools are to be developed which meet the user's needs, then we need to have a clearer understanding of how experienced practitioners achieve identification. We need to address questions such as: how do people narrow down a specimen to manageable group? and how do people select an appropriate key to use for an unknown specimen? Most importantly, do fieldworkers and specialised taxonomists approach identification in the same way and use the same characters? If they do not then tools developed for or by one group of users may be inappropriate for another group. We need to identify the strategies and shortcuts used in identification, and these should then be incorporated into new tools to make them more effective. If we are to address such issues, then those developing new identification tools need to work with psychologists, or knowledge engineers, and use knowledge acquisition techniques similar to those developed to assist with expert system development, in order to develop appropriate models on which to base species identification tools.

2.8 Are the tools effective?

Surprisingly little work has been done to evaluate the success of currently available identification tools, one exception is Tardivel and Morse (1998). We need to know which identification tools work and which do not. Evaluation of the tools must take place under the conditions in which they will be used, by the people who will be using them. Rigorous evaluation of tools will enable resources to be directed towards those tools which are shown to be most effective under different circumstances.

2.9 Understanding the identification process – conclusions

We need to know the answers to these questions to work out where efforts in developing species identification tools will be best directed. If, for example, it turns out that most identification is currently done in tropical or subtropical regions by minimally trained biologists assisted by parataxonomists, then tools are required which take into account the strong possibility of finding new species, which do not assume specialised biological knowledge, and which include a training element.

3. How do We Decide Which Technique to Use?

While there is still much we need to know about who does identification, where they do it and why, it is possible to offer limited guidance on which techniques are most suitable for a given species group. The choice of technique depends on: the tools which are already available; the characteristics of the group to be identified; and the objective of identification. Five groups of identification tools are considered: conventional paper keys; hypertext keys; multi-access keys; expert systems; and automated identification systems based on neural networks and multivariate statistical techniques.

This section draws on Edwards and Morse (1995) who reviewed computer-based species identification techniques and identified a number of questions surrounding the choice of an appropriate identification tool for a given species group.

3.1 What tools are currently available?

If any tools are available for the species group, are they appropriate for the task? For example, if a key is available, is it a good key? Some keys are notoriously difficult to use in that they are badly structured and have complex leads. The key must be appropriate in terms of both the geographical region covered, and in addressing the needs of the intended users. For example, AIDGAP keys are rigorously tested and developed to be used by relatively inexperienced users.

If the key is felt to be appropriate for the intended task, should it be converted to a hypertext key? The advantages of doing this are that it simplifies navigation of the key and improves access to descriptions, diagrams and glossary definitions (Tardivel and Morse, 1998), hence a key on a handheld computer may be easier to use in the field than a paper-based key. However, it is essential to realise that a bad paper key will make a bad hypertext key.

3.2 Is a data matrix available (or can one easily be constructed)?

If there is no suitable key available, then the type of data available needs be considered before deciding on the most appropriate technique. For example, if the data is available, or can easily be collected, in the form of a character × species matrix then a multi-access key can be constructed reasonably easily. Data matrices may be constructed using published descriptions, but if characters are not included in all descriptions, the matrix will have gaps, however this does not mean that a multi-access key cannot be used (see White, this volume) but its efficiency is reduced. The advantage of a multi-access key is that a number of well-developed software packages are available (including CABIKEY, INTKEY and PANKEY) and only the data matrix needs to be supplied to produce an identification tool.

If data is collected most naturally as a character × specimen matrix then an identification tool based on neural networks or multivariate statistics is possible. For such approaches to have a significant advantage over more conventional tools then the data must be collected at least “semi-automatically”. If the specimen has to be examined in detail or dissected to collect the data (even if some of that data is collected semi-automatically), it is questionable whether an automated approach is appropriate, because it is likely that the person preparing the specimen will have made a partial identification before the automated identification begins. If the identification has been narrowed down to a few species this information should be taken into account by the automated identification system, and may make the use of such a system unnecessary.

A disadvantage of automated identification techniques is that they are a black box approach and the identification skills of the users of such a system are not developed. However, a potential role for automated identification systems lies in monitoring studies, if the collection and identification of specimens could be done automatically by equipment in the field, the results could be collected and analysed as required.

The use of a character × specimen matrix is not restricted to automated identification techniques. For example, such a matrix can be used to produce a discriminant function equation to separate species. This equation is independent of the tool which produced it, and can be incorporated into other tools (e.g. conventional and multi-access keys).

3.3 No key, no data matrix, expert available?

If no key is available and there is insufficient information to develop a data matrix, but there is an expert on the group who is both available and willing to help with development of a tool, then development of an expert system may be considered. The disadvantage of developing an expert system is that it is very time-consuming to do properly because it is necessary to understand how the expert does identification both in terms of the characters used and the strategy adopted. If a successful expert system is developed for one species group, there is no guarantee that it will be possible to generalise the approach to other species groups as these may require different approaches to identification.

The advantage of using an expert system is that it has superior powers of explanation compared to other identification tools, and this will enhance its role in teaching and training. In addition, a well-constructed expert system will be the best tool to indicate new species and give useful incomplete identifications.

3.4 Final comments

The key to developing a successful identification tool is character selection. Hence, developing any identification tool involves working with experts in identification (these may be either specialised taxonomists or experienced practitioners) to identify the appropriate characters for a particular group and a particular group of users, and this requires that we have a deeper understanding of the identification process than is currently the case.

4. Conclusions

Three principle conclusions can be drawn:

A. We need answers to the questions outlined in section 2, in particular, we need to understand the identification process. Without these answers we are in danger of wasting resources on developing inappropriate tools.

B. There is a wide range of tools and techniques available. Some of these need further development (principally the automated techniques) others (mainly multi-access and hypertext keys) do not. Current multi-access key technology should be used to develop keys for appropriate species groups, whereas considerable research effort is needed to develop practical automated systems.

C. Given the diversity of both species groups and objectives of identification, it is highly unlikely that a single approach will be appropriate under all circumstances. We need to develop guidelines as to which tools are suitable in different situations, taking into account the taxonomic group, the skills and experience of those doing the identifications, and the circumstances under which identification is done.

5. Acknowledgements

Thanks are due to David Morse (University of Kent) for useful discussions on the questions raised in this paper.

6. References

Alberch, P. (1993) Museums, Collections and Biodiversity Inventories. *Trends in Ecology and Evolution* **8**, 372-375.

Chesmore, E.D., Femminella, O.P. and Swarbrick, M.D. (1998) Automated analysis of insect sounds using time-encoded signals and expert systems - a new method for species identification. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 273-287.

Dallwitz, M.J., Paine, T.A. and Zurcher, E.J. (1998) Interactive keys. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 201-212.

Edwards, M. and Morse, D.R. (1995) The potential for computer-aided identification in biodiversity research. *Trends in Ecology and Evolution* **10**, 153-158.

Gaston, K.J. and May, R.M. (1992) Taxonomy of taxonomists. *Nature* **356**, 281-282.

Tardivel, G.M. and Morse, D.R (1998) The role of the user in computer-based species identification. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R. (eds), *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 247-259.

REPORTS OF WORKING GROUPS

WORKING GROUP 1

AUTOMATED SYSTEMS

Chairman's Notes

Chairman: David Chesmore
Members: Mark O'Neill, Paul Bridge, Lynne Boddy, Colin Morris, Charles Lane

The general consensus among the group members was that automated systems are still very much in their infancy, with no actual commercial systems available. A number of points arose during discussions:

a. Applications. The main forms of application for automated systems are: detection of presence of taxa (e.g. pest detection), species discrimination and counting of individuals/species (e.g. biodiversity studies).

b. Identification of the Problem. It is vitally important to be able to adequately identify all aspects of the problem, for example, the methods for input (acoustics, images, analytical methods, etc.), classification (PCA, ANNs, expert systems, etc.), and to determine which are the most appropriate. The intended accuracy is also important to ascertain.

c. Validation. Since each system will have different inputs, feature extraction and identification techniques, it is important to be able to validate the different methods in ways that can be compared. For example, how can classification accuracies between different forms of ANN be realistically compared?

d. Interoperability. It is anticipated that many different systems will be integrated, creating a need of interoperability. Examples include common data formats, common core software and operating environments. Increasing use of the Internet may force this issue.

It was suggested that the following be carried out:

- develop a list of available techniques;
- create a database of skills;
- create a database of relevant publications;
- develop a research network of interested research groups;
- develop a Web site to hold the above information;
- identify a small number of demonstration systems that can be used to illustrate the viability of automated identification. Some of the systems described in this publication can be developed further for this purpose.

WORKING GROUP 2

KEY SYSTEMS

Chairman's Notes

Chairman: Bill Hominick
Members: Paul Kirk, Richard Pankhurst, Paul Beales

During the workshop, extensive discussions on electronic keys and databases took place. The four main recommendations, which are not restricted to nematodes, were:

1. It is not necessary to devote time or resources to developing new tools. Products demonstrated at this meeting showed that the tools available are adequate for present requirements.

2. Any product developed must be taxonomically sound, but it should be aimed at users who are not trained taxonomists.

3. A specific group of organisms should be chosen for a first key in a LOOP. It should emphasise the economically important species or genera. It must fulfil a demonstrated need. It should be visible on the Internet and its development will probably be facilitated by Internet links between collaborators. The product should be platform independent and should support training/educational needs.

4. There is no accepted standard database for biological/taxonomic uses. Therefore, before developing any database in BioNET-INTERNATIONAL LOOPs, the Technical Secretariat and EuroLOOP should be consulted for advice. This will:

    a. Avoid creating or duplicating a database that already exists.
    b. Prevent keyboarding information that is already available electronically.
    c. Ensure that it is compatible with others that may exist.

WORKING GROUP 3

EDUCATION AND TRAINING

Chairman's Notes

Chairman: Marion Edwards
Members: Chester Dent, Neil McAleece, David Morse

The objective of the workshop was to consider how computer-based tools could be used to provide support for education and training for individuals in developing countries whose job involves species identification, for example, those working in crop protection, plant quarantine and biodiversity studies. The outcome of the workshop was three principle recommendations:

I. The need for three surveys or reviews to establish: the needs of the target community; the currently available tools for education and training in species identification; and how computers are most effectively used in education and training.
II. A method to disseminate the above information.
III. A pilot project to demonstrate the potential of computer-based education and training tools to support those working on species identification in developing countries.

1. Surveys and Reviews

Before BIGCAT can establish guidelines for the provision of computer-based tools for education and training, there is a certain amount of ground work which needs to be done, to avoid reinventing the wheel and to ensure that the most effective tools are developed. Three areas of investigation were identified:

1.1 Establishing the needs

A survey is needed of the user community to determine:

I. Who is doing identification? The numbers of different groups of people doing identification (e.g. trained taxonomists, trained biologists and experienced fieldworkers, and untrained biologists and parataxonomists) must be determined.
II. What are they identifying? This needs to be done in terms of both the species groups being identified and the geographical regions where identification is taking place.
III. Why are they identifying it? This will influence the type of tool required, for example, is the objective to achieve an identification so that a treatment can be recommended (e.g. crop protection)?
IV. Where are they identifying it? This information is needed both in terms of geographical location, and in terms of laboratory and field-based identifications.
V. What technology is available for their use? Before developing computer-based tools, the target group's access to computers must be established, both in the laboratory and in the field. It is also essential to determine if appropriate technical support for the computers can be provided.
VI. What can they do already? It is necessary to establish the prior training of the target group, so that tools can be developed at an appropriate level. For example, parataxonomists may have completed secondary school but have no further education, clearly the tools developed must address this and assume a relatively low level of general biological knowledge.

Once these questions have been answered, the areas of greatest need for education and training tools can be determined.

1.2 Establishing what is available

In addition to establishing the needs of the users, it is necessary to establish what computer-based tools for education and training in species identification are already available. This would require a survey of the members of EuroLOOP who will either have or know of such tools. These tools need to be evaluated to determine:

a. What species groups they cover.
b. What geographical regions they cover.
c. For which groups of individuals they are suitable.

From this it will be possible to build a catalogue of available tools with guidance as to when they are most suitable. The survey should also establish which techniques used in the tools are most appropriate under different circumstances, and so give guidelines for developing new tools.

1.3 The effective use of computers in education

A third strand of the survey work involves reviewing what is known about education and learning, both in the context of computer-based learning (what are, and are not, appropriate techniques), and in terms of distance learning (much of the expertise lies within EuroLOOP and the need is in the other BioNET-INTERNATIONAL regions). It is also necessary to determine how much face-to-face training is needed for optimal use of the tools. It is recognised that locally based biologists and fieldworkers will have a key role in training those for whom they are responsible, but the need for support from those responsible for developing the tools must be determined.

The Open University has substantial experience in distance learning and is increasingly using computer-based approaches to education and it would be useful to draw on this and other appropriate expertise. There will be special challenges in this area with developing tools in appropriate languages.

It would also be hoped that this review could distinguish between the requirements of training tools and tools which provide on the job training. Conventional training tools would be useful to those starting in species identification, but on the job training has the advantage that the training is appropriate to the current task being undertaken.

1.4 Conclusion

These three areas initially suggest that there is a lot of work to be done. However, they would consist of independent surveys and reviews. In particular, the survey of the users should draw on the members of the existing BioNET-INTERNATIONAL LOOPs and would enable the needs of the BioNET-INTERNATIONAL communities to be established. The review of computer-based education and training should be done by someone active in the field of education rather than a biologist.

2. Dissemination of Information

The information which needs to be disseminated from the above surveys relates principally to the tools which are currently available, and the circumstances under which they are most appropriate. This type of information would be valuable to those requiring a tool to address a particular training need, either to see if one is available or to see which would be the most appropriate type of tool to develop. It is felt that this type of information could be effectively disseminated on a Website on the Internet.

3. A Pilot Project

The third area identified during the workshop would involve a substantial financial commitment. The aim would be to develop an appropriate educational or training tool. It was felt that the tool should be in use within the next five years, and in order to achieve this that the tool should be:

I. For use in a region with an appropriate infrastructure, in terms of computer support and preferably Internet access. The latter would simplify communication between the developers and users of the tool.
II. Based on existing, either published or unpublished, information rather than collecting primary data. However, it was recognised that using published information may have copyright implications.

It was also felt that in order to increase funding opportunities that the tool should address an economically important group. The tool should also be developed in conjunction with those responsible for training the intended users, as they are in the best position to understand local needs.

The stages in development would be:

i Identification of an appropriate identification problem.
ii Identification of an appropriate technique.
iii Development of the tool.
iv Evaluation of the tool under appropriate conditions.

Rigorous evaluation, both in terms of the effectiveness of the tool and the efficiency of development, is needed to demonstrate that the investment in time and money in the development of such tools is well spent. Assuming that the project was successful, it would help pave the way for the development of future tools.

4. Conclusions

Members of the workshop felt that until the survey work outlined in the first section was completed, it was difficult to make firm recommendations on the development of computer-based tools for training and education in species identification. However, it was also felt that the early development of a training tool was essential to show that the approach was both practical and valuable.

WORKING GROUP 4

IDENTIFYING THE OVERALL NEEDS OF BioNET-INTERNATIONAL WITHIN THE CONTEXT OF COMPUTER-AIDED TAXONOMY

Chairman's Notes

Chairman: Bill Parker
Members: John Lambshead, David Minter, Simon Gallagher

The initial group consensus was that the question of 'needs' was fundamental to the deliberations of the other Working Groups as well as cutting across the four principal BioNET-INTERNATIONAL programmes. Thus it would be necessary to take a very broad brush approach to assessing needs to ensure that the conclusions would be generally valid across the whole spectrum of BioNET-INTERNATIONAL's operation.

Two fundamental topic areas were addressed:

1. What do BioNET-INTERNATIONAL 'users' need and why?
2. What are the main strands inherent in the taxonomy/biodiversity issue?

What do BioNET-INTERNATIONAL users need and why?

Three principal groups of users were defined:

User groups	Who	What	Objectives
LOOPs	e.g. SAFRINET	e.g. digitising IMI descriptions of pathogenic fungi & bacteria	Met local need, but objective ill-defined
Funders	Governments, ODA, World Bank, UNDP etc.	Specific programmes proposed by LOOPs or other bodies	Meeting obligations under Rio Convention
Governments	All	As above	As above

It was clear from this analysis that funding bodies and governments were meeting their (mainly political) objectives, but that fundamentally, the BioNET-INTERNATIONAL LOOPs had not adequately defined their own local needs and objectives, or if they had these had not been communicated back to the BioNET-INTERNATIONAL Technical Secretariat. For BioNET-INTERNATIONAL to function effectively, it was therefore necessary to do a 'market research' type survey to ascertain the needs of the LOOPs so that these could be adequately addressed within the BioNET-INTERNATIONAL core programmes.

What are the main strands inherent in the taxonomy/biodiversity issue?

The view was strongly put forward that there were two fundamental issues associated with managing and accessing taxonomic/biodiversity data:

1. Managing and accessing the mass of paper-based information (referred to as 'meta-data' in the discussions) associated with specimens 'locked-up' in institutions and museums around the world. This largely untapped source of data probably contains much of the data that is required to provide the LOOPs with fundamental taxonomic data relevant to their regions, but it was difficult to access by the local community let alone the wider one.
2. Coping with the collection of new specimens and associated data. The Group discussions suggested that collecting more data was only making an already difficult data handling situation worse. However, in the report-back session the view was expressed that in some circumstances, it could be cheaper for developing countries to go out and make new collections and electronically catalogue them correctly than it would to wade through existing collections to access the same data.

Accessing and managing paper-based data

Of the two areas, it was felt that managing the existing paper information remained a major priority. Associated with this were a number of key areas and processes which required attention. These were:

1. Developing a 'central names database' (possibly already in existence for some phyla) which would act as the central reference point for taxonomists and would ensure that everyone used standard nomenclature.
2. Ensuring standardisation of geographic referencing of specimen collection location. This would ease, for example, the standard use of data in geographic information systems (GIS). This point is more critical than it might appear at first sight, as if the paper data is electronically referenced by location as well as by phylogeny, the task of extracting data relevant to particular countries (or LOOPs) becomes infinitely easier (see also below).

Both these problems are technically solvable and are pre-requisites to the next stages involved in dealing with the paper-based data resource. These stages are:

1. Getting the paper-based data into an electronic database, preferably in a standard (or easily convertible) format.
2. Validating the databases (i.e. ensuring correct data entry).
3. Identifying 'holes' in the data - to avoid 're-inventing the wheel'.
4. Filling holes in the data (including by new collection).
5. Making the data accessible to potential users.

There are a number of problems associated with this approach.

1. Most of the paper-based data is in the First World and the former Eastern Bloc countries, but it is not clear exactly where it is (EuroLOOP may have a role here).
2. The paper-based data resource is so vast and so inaccessible that it is simply not possible to contemplate electronically cataloguing it all.
3. The data is largely not quality-controlled, and therefore lacks credibility.
4. It is essential to ensure easy transfer of data around the world that electronic standards for software and data entry are set. There was support for BioNET-INTERNATIONAL to organise a workshop on this topic.

These problems mean that it is essential to prioritise what is to be extracted from the existing paper data resource. The primary way to set priorities should be to identify the needs of the LOOPs and let these define which data should be extracted, possibly on a country-by-country basis (hence the need for correct geographical referencing, see above). This links directly back to the need for ‘market research’ within the LOOPs identified at the start of the discussions (see above). Once this market research has been done, individual projects can start to be formulated by the LOOPs which address their particular needs, and which should also meet the political objectives of the funding ‘customer’. Any projects completed should be formally evaluated by BioNET-INTERNATIONAL (if evaluation is not already present in the project management/reporting structure) to ensure that the original objectives have been properly met.

Summary of Needs

1. Market research with LOOPs to identify what their needs are at a local level.
2. Development of a central names database.
3. Standardisation of geographical referencing.
4. Identify where the paper-based data resource is actually located.
5. Standardise electronic data input - possibly by running a workshop.

THEN

Bring forward project proposals from LOOPs which meet their local requirements (based on market research) and which could either:

a. draw on the existing paper-based resource, or
b. involve a carefully targeted amount of new collection and immediate electronic cataloguing.

Ancillary Problems

The group also identified two other issues which BioNET-INTERNATIONAL needs to consider:

1. Avoidance of duplication of effort with other organisations.
2. Identify who and where the current taxonomic expertise resides (there is considerable expertise in the former Eastern Bloc which is largely unknown to the general taxonomic community at present).

Published by the BioNET-INTERNATIONAL
Technical Secretariat with the financial assistance of the
United Kingdom Department for International Development (DFID).