Zootaxa 2217: 43–55 (2009)
www.mapress.com/zootaxa/

Forum

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

Accelerating taxonomic discovery through automated character extraction

JOHN LA SALLE[1], QUENTIN WHEELER[2], PAUL JACKWAY[3], SHAUN WINTERTON[4], DONALD HOBERN[5], DAVID LOVELL[6]

[1] *CSIRO Entomology, GPO Box 1700, Canberra, ACT, 2601, Australia; E-mail: john.lasalle@csiro.au*

[2] *International Institute for Species Exploration, School of Life Sciences, Arizona State University, PO Box 876505, Tempe, Arizona, 85287-6505, USA; E-mail: Quentin.Wheeler@asu.edu*

[3] *CSIRO Mathematical and Information Sciences, Queensland Bioscience Precinct, 306 Carmody Road, St Lucia, Queensland, 4067, Australia; E-mail: Paul.Jackway@csiro.au*

[4] *University of Queensland Insect Collection, School of Biological Science, University of Queensland, St Lucia, Queensland, 4072, Australia; E-mail: wintertonshaun@gmail.com*

[5] *Atlas of Living Australia, CSIRO Entomology, GPO Box 1700, Canberra, ACT, 2601, Australia; E-mail: Donald.Hobern@csiro.au*

[6] *CSIRO Mathematical and Information Sciences, GPO Box 664, ACT, 2601, Australia; E-mail: David.Lovell@csiro.au*

Abstract

This paper discusses the following key messages. Taxonomy is (and taxonomists are) more important than ever in times of global change. Taxonomic endeavour is not occurring fast enough: in 250 years since the creation of the Linnean *Systema Naturae*, only about 20% of Earth's species have been named. We need fundamental changes to the taxonomic process and paradigm to increase taxonomic productivity by orders of magnitude. Currently, taxonomic productivity is limited principally by the rate at which we capture and manage morphological information to enable species discovery. Many recent (and welcomed) initiatives in managing and delivering biodiversity information and accelerating the taxonomic process do not address this bottleneck. Development of computational image analysis and feature extraction methods is a crucial missing capacity needed to enable taxonomists to overcome the taxonomic impediment in a meaningful time frame.

Key words: taxonomy, taxonomic impediment, automated character extraction, image analysis, feature extraction, pattern recognition

Introduction

There is nothing new about taxonomy: it is the oldest of all the sciences. However, there is a growing feeling that taxonomy is now more important than ever, particularly as we need to understand enough about ecosystem function to make informed natural resource management decisions in an era of global change. There is clearly a need to shift taxonomic endeavour into the digital era to improve the pace at which we can supply taxonomic products and information. Recent volumes and papers have stressed these points, and offered compelling suggestions about how to accelerate taxonomic productivity (Godfray 2002a,b; Godfray & Knapp 2004; Wilson 2004; MacLeod 2007; Wheeler 2008).

These works recommend critical steps forward for improving the taxonomic process, and the pace of species description has increased dramatically as a result. Roughly 1.8 million species have been described in 250 years, giving an average rate of about 7200 species/year. Today we average somewhere between 16,000 and 20,000 per year which is more than double the historical average (SOS Report 2009). This is a clear indication that taxonomists are cognizant of the need to increase the rate of species discovery and description, and are working towards that goal.

Accepted by Z.-Q. Zhang: 31 Jul. 2009; published: 2 Sept. 2009

Although this increase in taxonomic productivity is encouraging, as yet the real problem limiting taxonomic productivity has not been addressed: *taxonomists are still gathering morphological information and describing species as they have done for the last 250 years*. While increased funds for educating and supporting taxon experts is essential, so too are strategic investments in technology that can add efficiencies and speed to their work.

With tremendous undescribed biodiversity on Earth, the societal need of taxonomy is greater now than ever, yet resources supporting taxonomy are becoming scarcer (Wheeler *et al*. 2004). This is the *taxonomic impediment*, and simply means that despite identifying the problem, we still lack the practical taxonomic means to describe the remaining biodiversity on Earth (Evenhuis 2007).

To address this challenge, we suggest meaningful and innovative change in the taxonomic process by automating it to a much greater extent. This encompasses digital phenotype capture, visualization, analysis, search and retrieval. It would involve image library creation, morphological feature extraction, application of standard descriptive ontologies, and statistical and database analysis to automate the delineation of species, and create species descriptions, revision, keys, hypotheses of evolution and other taxonomic outputs.

The real paradigm shift that we are proposing in the taxonomic process is the development of computational methods to extract morphological information from images. The advances in phenotypic visualization and feature extraction necessary to support accelerated species discovery and description will need to be developed by specialists in image analysis, statistics, computer science and software engineering, working in conjunction with taxonomists and biodiversity informaticians. Only through such radical change in the taxonomic process can we hope to document life on Earth in a time frame that might enable us to manage and preserve it.

The goal is not to replace taxonomists but rather to increase their output by orders of magnitude. Taxonomists (and proper taxonomic methodology) will be an integral part of any process of discovering and describing life on earth. Taxonomy is not a mindless, pragmatic enterprise that can be fully automated, and evolutionarily informative characters are not the same as phenotypic correspondence. In this regard, the proposed system of automated character extraction should not be confused with phenetic exercises: it is the extraction of a suite of characters that have been chosen by a taxonomist as having evolutionary/phylogenetic value. In-depth knowledge of taxa and skill in comparative morphology will be requisite to forming hypotheses about sameness among ancestral and subsequently modified phenotypes. Only once such hypotheses about homology exist will it be possible to create correspondence maps and ontologies that a computer could follow to place a correct value in a matrix. Our proposition is to develop new and improved tools and systems to accelerate taxonomists' ability to characterise, compare and contrast living organisms, dramatically augmenting their ability to deliver new insights. It is important to remember that well-supported hypotheses of evolution will allow us to create classifications whose predictive value will be of critical importance to scientists trying to understand how organisms might respond to global change.

The taxonomic impediment

What is the taxonomic impediment?

The Convention on Biological Diversity acknowledged a "taxonomic impediment" which prevents other biodiversity research and blocks biodiversity outcomes, and builds on previous arguments presented by Wilson (1985). Effective biodiversity management depends on taxonomic knowledge, and there is a lack of taxonomic knowledge that prevents other biodiversity research and thus impedes biodiversity outcomes. This becomes particularly significant when the species we don't yet know may be the most important in ecosystem function**.** Natural resource management depends on critical ecosystem services provided by lesser known groups such as invertebrates and microorganisms (e.g., pollination, nitrogen fixation, decomposition, soil conditioning).

The ability to distinguish and identify organisms is not only relevant to biodiversity. It is equally essential to enabling research and applied outcomes in activities such as biosecurity, natural resource management, biodiscovery, predictive modelling, policy, and evolutionary biology. In a world suffering the effects of global change, the societal need for taxonomy has become more important than ever.

How should we respond to the Taxonomic Impediment?

We must have clear thinking on how we respond to the taxonomic impediment and identify its root causes. Unfortunately, some taxonomists seem to see it merely as a justification to do what they have always done. We see it as a challenge that demands an active response, and a call to adapt our thinking to the problem that needs to be solved, rather than using the problem to support our thinking. As such, we should set some tough goals for ourselves and our science. One such goal would be to describe all life in the next 50 years. A strategy to accomplish this must include further funding for taxonomy combined with the development of new technologies to help accelerate the process.

This paper proposes that we will gain the necessary "orders of magnitude" shift in taxonomic productivity by applying computational methods to phenotypic analysis to automate character extraction. This will be done by developing automated or semi-automated methods to analyse an image of an organism, automatically recognize pre-determined morphological characters, and populate a data matrix which can then be used to produce diagnoses, descriptions, diagnostic tools and hypotheses of evolutionary relationships. This new generation of tools will only have the necessary impact if it is placed in the hands of a properly trained and staffed workforce of taxonomists.

Modern initiatives in taxonomy and biodiversity information management

There are a variety of initiatives which are providing a foundation for the new taxonomy by facilitating and accelerating the taxonomic process and the managing and delivery of biodiversity data. All of these initiatives are to be praised for their vision and achievements, and all will be integral components of any cohesive system that must be developed to overcome the taxonomic impediment. The following is a very brief overview of some of these.

Managing and delivering biodiversity data

Tremendous advances have been made in the fields of managing and delivering biodiversity data, and more importantly, in forming networks and collaborations to promote these activities. There are an increasing number of global scale initiatives that are intended to make biodiversity data freely available to a diverse user community for a variety of purposes. A few of the more notable of these include:

- Encyclopedia of Life (EoL)—http://www.eol.org
- Global Biodiversity Information Facility (GBIF)—http://www.gbif.org
- GenBank—http://www.ncbi.nlm.nih.gov/Genbank
- Atlas of Living Australia (ALA)—http://www.ala.org.au
- Species 2000—http://www.sp2000.org
- ITIS Catalogue of Life—http://www.catalogueoflife.org
- MorphBank—http://www.morphbank.net
- ZooBank is the official online registry for Zoological Nomenclature—http://www.zoobank.org
- Biodiversity Heritage Library (BHL)—http://www.biodiversitylibrary.org

These initiatives have had a tremendous impact on taxonomy, and are producing positive changes by setting new standards for collaboration and resource sharing, global thinking, and delivery to a wide range of stakeholders. However, it must be remembered that while these initiatives all contribute to efficient and accelerated management and delivery of information, the real bottleneck is the production of biodiversity information in the first place.

Accelerating taxonomic research

One of the central premises of this paper is that the pace of taxonomic endeavour is not sufficient to the task. This is not a new idea. Several authors have treated this recently, and there are a variety of new initiatives aimed at accelerating taxonomic research. Wheeler (2007, 2008) mentioned recent funding initiatives from the US National Science Foundation: Partnerships to Enhance Expertise in Taxonomy (PEET), Assembling the Tree of Life (AToL), Revisionary Syntheses in Systematics (RevSys) and Planetary Biodiversity Inventory (PBI). Similar projects aimed at increasing taxonomic productivity using eResearch thinking have been funded in Europe (EDIT, European Distributed Institute of Taxonomy, http://www.e-taxonomy.eu; CATE, Creating a Taxonomic e-Science, http://www.cate-project.org) and Australia (TRIN, Taxonomic Research and Information Network, http://www.taxonomy.org.au).

These initiatives have several themes in common: increased networking and collaboration, coordination of expertise, knowledge sharing, and the creation of virtual workspaces. In particular, Wheeler (2007) discusses cybertaxonomy as an emerging field that will utilise digital technology, information science and computer engineering to bring experts and information together into knowledge communities to substantially increase the quality and quantity of taxonomic output.

New paradigms for publication

Describing species using traditional methods is very slow and tedious by nature, and, considering estimates of the number of new species requiring description, radical changes are needed to this taxonomic process if we are to have any chance of documenting biodiversity in a realistic timeframe. Advances on how we handle information in the digital era are producing paradigm shifts in publications. The journal *Zootaxa* has grown to mega-journal status by combining rapid publication times, affordability, and making all papers available through the web (Zhang 2006, 2008b). In addition, the electronic format of the journal's web-based papers allows for the utilization of embedded links and a variety of cybertaxonomic tools to enhance the quality and utility of publications. This ability has recently been showcased in a publication on *Chromis* fishes (Pyle et al. 2008), which sets a higher standard for the next generation of taxonomic publications. Zhang (2008a) pointed out some of the innovations in this paper, which include:

- New species with their scientific names prospectively registered in the official ZooBank registry developed by the International Commission on Zoological Nomenclature (http://www.iczn.org)
- Descriptive data marked up with SDD (TDWG standard for descriptive data http://www.tdwg.org/activities/sdd), to enable direct downloading of raw data
- Species descriptions marked up with XML tags using standards in TaxonX (http://research.amnh.org/informatics/taxlit/schemas) and taXMLit11 (http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf)
- Images with embedded links to images deposited in the MorphBank (http://www.morphbank.net)
- Cited specimens with embedded links to online databases in museums and/or via the GBIF portal (http://data.gbif.org)
- DNA Barcodes deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank), in compliance with the Barcode of Life Data Systems (BOLD, http://www.barcodinglife.org) and associated Fish Barcode of Life Initiative (http://www.fishbol.org)
- References cited registered with ZooBank, with some available as full-page images through the Biodiversity Heritage Library (BHL, http://www.biodiversitylibrary.org).

This procedure of value-adding to taxonomic papers by Pyle et al. (2008) was followed by Deans & Kawada (2008) and Johnson *et al*. (2008). Expanding upon these authors, Winterton (2009) recently published a revision of the Australian therevid genus *Neodialineura* also embedding numerous web resources in the document, and using a character state matrix to produce natural language descriptions for use both in original descriptions and in an online interactive key. This test case revision was significant because it was produced in a highly standardised way, and in approximately one-third of the time normally taken to publish a taxonomic revision of the equivalent size. The success of *Zootaxa,* and the advances in cybertaxonomy shown by the Pyle et al (2008), make even more compelling the case for making electronically published names available under the codes of nomenclature. Indeed, this subject is now under consideration by the International Code of Zoological Nomenclature (International Commission on Zoological Nomenclature 2008). There is little doubt that web-based applications, products and publications are the way of the future for systematics.

Automating taxon identification

Automating the identification of species is not a new concept, and computer assisted identification systems have been in existence for over two and half decades (e.g., Daly et al. 1982), with considerable progress having been made since that time. An argument for the implementation of automated species identification systems was offered by Gaston & O'Neill (2004), and an edited volume on the subject presents an overview of the current state of knowledge in the field (MacLeod 2007). Systems that are currently operational and which have shown some success in automated identification include: DAISY (Digital Automated identification System), ABIS (Automated Bee Identification System), SPIDA (Species Identification Automated). These are reviewed in MacLeod (2007). In addition we can add: Automated Leafhopper Identification System (ALIS), and Automatic Identification and characterization of Microbial PopulationS (AIMS). Gaston & O'Neill (2004) give the following summary table of systems and technologies:

TABLE 1. Some automated taxonomic identification systems based on morphological characteristics. From: Gaston & O'Neill (2004).

Name	Method	Reference
ALIS Automated Leafhopper Identification System	Discriminant function	Dietrich & Pooley (1994)
DAISY Digital Automated Identification (SYstem)	Lucas continuous *n*-tuple classifier/ PSOM network	Gauld *et al.* (2000), O'Neill *et al.* (2000)
AIMS Automatic Identification and characterization of Microbial PopulationS	Artificial neural networks (ANNs)	Jonker *et al.* (2000)
ABIS Automatic Bee Identification System	Support vector machines, kernel discriminant analysis	Arbuckle *et al.* (2001); Arbuckle (2002)

Most of these systems require the removal of "noisy" backgrounds from images, and accurately controlled pose, lighting, cropping, and scale of the specimens in the images. Furthermore these aspects often have to be performed manually by the operator and, if not done perfectly, performance suffers. To fully benefit from the power and speed of computer pattern extraction, attention needs to be directed to increasing the robustness of systems and automatically handling these practical issues.

Automating taxon identification will be an important component of any future system that might fundamentally change systematics, but it will not serve the same purpose as is intended for using feature extraction for phenotypic visualization and analysis. Gaston & O'Neill pointed out that automated species identification should make a "valuable contribution to reducing the burden of routine identifications" (Gaston & O'Neill 2004: 655), and we agree wholeheartedly with this statement. The compelling need in taxonomy, however, is discovering and describing the species we do not know, and not just having the ability to accurately identify from a library of species that we do know. To this end, it is necessary to build upon the framework provided through automated taxon identification to have real impact on the taxonomic impediment.

Integration of molecular data

It is clear that the integration of molecular data with morphological taxonomy will benefit efforts to describe and understand life. Proponents of DNA barcoding have suggested it is a method that could use short, standardized gene regions to automate species identifications, as well as providing hypotheses as to presence of cryptic species and overall evolutionary relationships (Hebert *et al*. 2003; Tautz *et al*. 2003; Hebert & Gregory 2005). Other authors feel that molecular taxonomists can best contribute to taxonomy by taking a more integrative approach that would combine the use of potentially a variety of genes with other areas of taxonomic effort (Will *et al*. 2005; Dayrat 2005). Both of these approaches hold benefits to the taxonomic community, although molecular methods on their own can not replace taxonomists. Molecular data should complement, but never replace, taxonomic endeavor.

In contrast to the pace at which taxonomic data is generated, there is currently an explosion in the volume and variety of molecular data thanks to a range of new “omics” measurement platforms. By creating a matching explosion of phenotypic data we increase the potential for a dramatic increase in the understanding of the relationship between genotype and phenotype. Automated species discovery stands to deliver benefits to molecular biology as well as taxonomy.

The big gap

The initiatives are building on developments in digital technology, information science and computer engineering for better management of existing information, and for improving our networking and collaboration skills. As yet, however, we are not attempting to apply technological advances to automate the population of the morphological data matrix which will support a variety of other taxonomic activities and products.

Creating the capacity to automate aspects of phenotypic visualization and analysis will provide the most significant tool which is currently missing in our bid to overcome the taxonomic impediment.

Taxonomy tomorrow

Automating aspects of the extraction and recording of morphological information, and placing this information in a character data matrix, has the ability to transform the taxonomic process to enable meaningful progress against the taxonomic impediment.

There seems to have been a move away from populating morphological character matrices towards automated taxon classification which has, too often, bypassed extracting morphological characters and gone straight to classifying images on the basis of pixel intensities and arrangements. This has been driven, in part, by the desire to force taxon identification into the existing pattern recognition paradigm rather than the reverse. The result has been systems overly sensitive to image variation, operating in very narrow and controlled niches, and with adequate performance only after costly, labour- and data-intensive training phases. Being able to efficiently obtain an accurate character matrix from a specimen will not work in the absence of a trained taxonomic workforce, and it will work optimally when combined with the other modern initiatives discussed in the previous section. We need to automate every step of populating the biodiversity data matrix, and present all the above as an integrated package without any “stand alone” components.

The taxonomic process of the future could be as follows:

Creation of high quality image libraries

Technicians will be able to scan specimens to automatically produce high quality images. In the future, this might include a wide range of imaging technologies, such as photographs, 3D imagery, holograms, X-rays, MRIs and CAT scans. This image library will be the foundation for other activities. Following on to the principles of other biodiversity information, the custodianship of the images would remain with individual institutions who would make them freely available through distributed databases. We note that one of the most time consuming steps in image acquisition is in preparing or mounting the specimen for imaging. With the steadily decreasing cost of imaging and data storage it will make sense to capture and permanently record high quality and high resolution images of the specimen from many angles and across many wavelengths and modalities.

Application of standardised morphological ontologies

A common, systematic and structured language of descriptive morphology is essential for modern cybertaxonomy. Standard terminology is especially important as a basis for automated systems and to enable atomised character data to be distributed in character databases. Specialist knowledge will be required define character states, determine homologies, and ready the databases for automated population.

Automated image analysis to extract a determined character set

Automated (or semi-automated) image analysis is an essential step in turning raw pixel data into morphological characters. This step presents a myriad of challenges as the ways in which humans and machines analyse image data differ significantly. It may be that some morphological features obvious to the human eye are opaque to machine algorithms, and vice versa. It may also be possible to exploit the human visual system's ability to detect changes in images in conjunction with the capacity of computers to rapidly collate and present large numbers of similar images.

The extraction and analysis of information from digital images has been an active research area for over four decades. We can repeatedly, robustly, and accurately, segment foreground from background objects in digital images (Cheng *et al.* 2001), detect and find distinctive image features (Lowe 2004), measure the colour (Klinker *et al.* 1990), shape and size (Rohlf & Bookstein 2003), texture of objects (Reed & Hans du Buf 1993), and their relationship to each other (Eshera & Fu 1986). The challenge is now to systematically apply these advances and understanding to recognize and extract morphological characters from raw images to form a numerical description (a "character matrix") of that specimen.

We foresee an extensible framework consisting of an array of character extraction modules operating on sets of images of the specimen, see Figure 1. Each module has been designed to robustly extract measure or compute a single morphological character (or small set of related characters). New characters are added by plugging new modules into the system. Over time as technology improves, improved modules replace older modules. With a suitable framework and standards in place to ensure compatibility this can become a shared international collaborative effort much as the software community has done with packages for *TeX* (www.ctan.org), *R* (cran.r-project.org*)* or *Perl* (www.cpan.org).

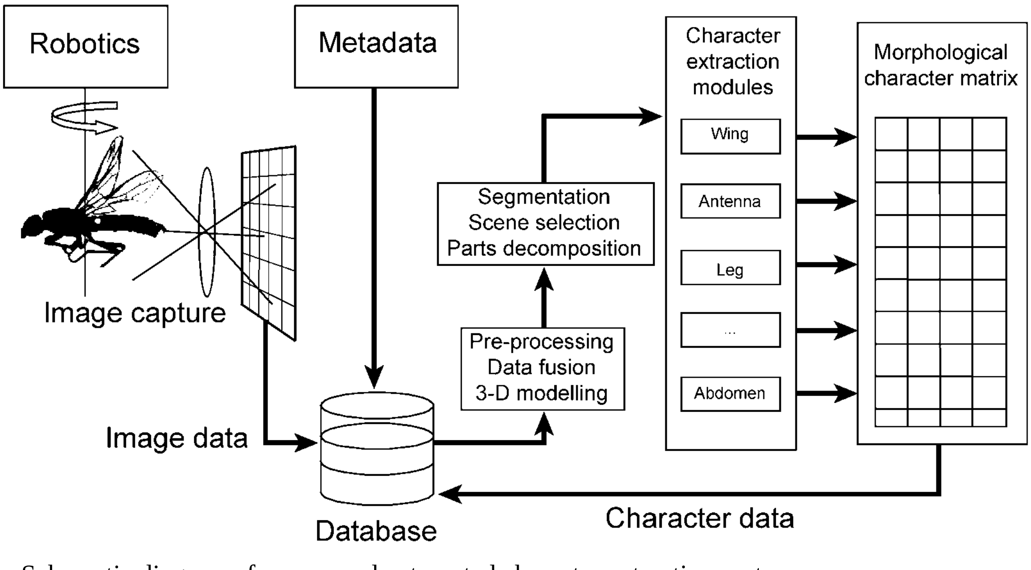

FIGURE 1. Schematic diagram of a proposed automated character extraction system.

Considerable research effort will be required to construct a system capable of industrial-scale capture of suites of phenotypic information, and algorithms to extract different types of character will have to be developed individually. In a simple example, Table 2 illustrates the steps in the process of extracting characters from insect wings, a simpler proposition than many other types of characters because they can be extracted effectively from 2D images.

TABLE 2. Steps involved in extracting insect wing venation characters.

Step	Purpose	References
Compilation of Image Database	Capture of the high quality images to subject to analysis.	
Preprocessing	Noise reduction, artifact rejection, data fusion, 3-D modelling	Besl & McKay 1992, Gonzalez & Woods 2007, Liu *et al.* 2005, Chow & Chan 2009
Segmentation and Parts Decomposition	Separate the insect from the background. Separate the wing from everything else in the image. This step ensures that only the wing is being measured.	Cheng *et al.* 2001
Feature Extraction and Pattern Recognition	Recognize the veins in the wing. This step will recognize and identify the different veins in the wing.	Houle *et al.* 2003, Tofilski 2004, Fedor *et al.* 2009
Character Computation	Extract characters of interest as defined in an ontology. These could include colour, texture, sizes, shape, ratios, angles and relationships of different characters to each other.	Eshera & Fu 1986, Klinker *et al.* 1990, Reed & Hans du Buf 1993, Rohlf & Bookstein 2003, Lowe 2004
Population of Character Database	Information from the extraction step feeds directly into the character matrix.	

The problem of homology

A critical step in the proposed process will be to translate hypotheses about characters into terms that the software can understand, i.e., to "train" the image analysis system to recognize concepts concerning homology and not merely phenetic similarity. The forewing of a dragonfly and beetle would hardly be recognizable by shape alone as the same character. There are many instances where the informative shape of parts of the aedeagus among species in a single genus are so different that they would not be easily recognized as same, and instances where non-homologous parts are so convergent that they might well be mistaken. Considering how hard it has been to arrive at database standards, it is not a trivial challenge to convert our conceptual character hypotheses into a sufficient "homology library" of images that define for the computer what is within or without the theoretical boundaries of a character. However, when this is achieved a computer could inform character assignments, rather than simply matching overall phenotype similarity.

In terms of a research pipeline, the first steps in automating character extraction must be through the process of extracting previously identified and agreed characters, with determinations as to homology and diagnostic value of characters made by a trained taxonomist familiar with a specific group. Creating such a system will open up numerous other research areas, with training computers to homologize characters being one such particularly interesting and challenging field.

Sorting of putative taxa

Statistical, machine learning and other pattern recognition tools will be used to rapidly assess morphological characters selected by taxonomists for sorting specimens into putative species, identifying named species, and flagging potentially novel species. This can be done through a combination of existing automated taxon identification tools as well as the comparison of the data matrices produced through feature extraction.

Accessibility of data

All data will be made freely accessible through the web (through initiatives such as GBIF, The Encyclopedia of Life, *Atlas of Living Australia*) to contribute to an on-line, dynamic biodiversity knowledge bank. As with many other forms of information, taxonomic information gains value the more widely it is shared, integrated and used.

Taxonomic products and tools

Specimen character matrices will be used to automatically produce descriptions, diagnoses and identification tools (keys, field guides, revisions, monographs), as well as being available for phylogenetic analysis. Furthermore, automated image capture affords much greater opportunities to characterise phenotype variation within species. Taxonomists will be able to start projects with a manuscript ready to review and refine rather than years of research to undertake. If done properly, taxonomists will start projects with a list of putative species, descriptions, illustrations, hypotheses of variation, keys, classification and phylogenies. Imagine being able to produce a monograph treating a few hundred species in a matter of months, with revisions being web-based and updates produced automatically.

Dynamic taxonomy

Not only is a small proportion of species known to science, those that have been described are often poorly known. For example, over half of beetles are known from a single locality, with 13% known from only one specimen (Stork 1999). Descriptions of species boundaries will be web-based, and dynamically updated with digitisation of new specimens which add to our understanding of phenotypic variability within a species. Taxonomic products will incorporate real-time accrual of changes and improvements, including new species as they are discovered and described.

Integration of data sets

Species names will be repositories for further information on the species, including biology, ecology, distribution, and trophic associations. Morphological information can be fully integrated with molecular data increasing our ability to explore and understand relationships between genotype and phenotype.

Automating the capture and integration of associated data

The automation of other forms of data capture (e.g., GIS data, satellite data, remote sensing, “omics” data) will contribute to accelerating the population of the knowledge bank with information of relevance to other endeavours.

Through this transformed process, which is intended to complement, rather than replace, other initiatives aimed at accelerating taxonomic research and delivery, taxonomists will achieve an order of magnitude increase in their productivity.

More importantly, other users of taxonomic products and information will achieve the same benefits as the taxonomic community. They will:

- Have a vastly increased biodiversity knowledge bank for informing a variety of policy, management and research activities.
- Work in a dynamic, virtual environment.
- Be able to produce descriptions, diagnoses, keys and field guides with the touch of a button.
- Have large data sets at their disposal so that they can explore higher level questions in evolutionary biology .
- Have access to other data types associated with any given species
- Have data linked to analysis tools.

What will be left to do?

Simply taking pictures of existing museum specimens, no matter how sophisticated the process, will not solve the entire problem. As one example, May (2004) pointed out that actually finding and collecting specimens in the field will be the bottleneck that limits the rate of taxonomic achievement. Other steps in the process of discovering and naming all life on the planet, such as sorting, mounting and processing, will also need attention, and automation where possible. These are rich areas to search for technologies to complement those that we are proposing in this paper.

Indeed, first generation technology has already been developed to automate the sorting of field generated samples. RAPID (Robotic Automated Pest ID) technology includes robotics, automated sample feeding, image analysis, and relational databases (Walters *et al*. 2008; Drake 2009).

Time for a reality check

What is being proposed in this paper is a radical solution involving technologies that have not yet been developed (and it should be pointed out that if it was easy, it would already have been done). The question remains as to whether we actually can develop the ability to automatically extract morphological character data from images of specimens. The simple answer is that *we have to*.

We have probably named between 1.5 and 1.8 million species, and current best estimates are that there are somewhere around 10 million species living on this planet (Wilson 2004). Given that the bulk of taxonomic endeavour has occurred in the last 150 to 180 years, we can assume a historic rate of description over the last century and a half of about 1 million species / 100 years. Initiatives discussed above are already increasing the rate of taxonomic productivity; however, even doubling our pace would still require several centuries to complete the inventory of life on earth.

We simply can not continue taxonomy at the current pace and believe that we are delivering the greatest benefit from our science. It is only through orders of magnitude increase in the discovery and description of life on this planet that we will have a chance of saving and managing the environment while faced with global change.

Despite the number of technical difficulties to be overcome, our best chance for the timely description of life on earth is by providing a trained taxonomic work force with proper levels of funding and support, combined with new technologies developed specifically to accelerate as many aspects of the taxonomic process as possible.

Who will be involved?

The taxonomic community cannot hope to create or implement the entire range of technological advances necessary to automated character extraction by itself. It must engage with experts in other areas to ensure that it is developing and using state of the art techniques. We envision that we would need to involve experts in the fields of

- Image capture
- Image Analysis (especially feature extraction) .
- Statistics and machine learning (especially multivariate statistics and classification).
- Linguistics (e.g. syntactic pattern recognition) .
- Information and Communication Technology (including software engineering and architecture)
- Biodiversity Informatics
- 3D image generation.

The real goals

Naming all species on the planet is a noble goal in itself. However, one of the follow-on benefits of naming all species is that it would provide the necessary starting point for halting biodiversity decline. A larger goal is that our great-great-grandchildren will inherit the same planet with the same rich biological diversity that we enjoy today. Increasing the rate of species discovery will create a knowledge bank to hold critical information about species, their biology, function and dependencies which will inform: policy, resource management, biosecurity, predictive modelling, automated identification, biodiscovery, evolutionary biology, and predictive classification. This will ensure relevance for taxonomy and biological collections, and help meet applied outcomes.

Acknowledgments

We would like to thank Jeff Drake, USDA, for supplying information concerning automated sample sorting. Anne Hastings, CSIRO Entomology, helped with the preparation of the figure.

References

Arbuckle, T. (2002) Automatic identification of bees' species from images of their wings. In *Proc. 9th Int. Workshop on Systems, Signals and Image Processing*, pp. 509–511. Manchester, UMIST.

Arbuckle, T., Schroder, S., Steinhage, V. & Wittmann, D. (2001) Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In *Proc. 15th Int. Symp. Informatics for Environmental Protection*, ETH Zurich, 10–12 October 2001, vol. 1, pp. 425–430. Zurich: Metropolis.

Besl, P.J. & McKay, N.D. (1992) A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 239–256.

Cheng, HD., Jiang, XH., Sun, Y. & Wang, JL. (2001) Color Image Segmentation: Advances and Prospects. *Pattern recognition*, 34(12), 2259–2281.

Chow, S.K. & Chan, K.L. (2009) Reconstruction of photorealistic 3D model of ceramic artefacts for interactive virtual exhibition. *Journal of Cultural Heritage*, 10, 161–173.

Daly, H.V., Hoelmer, K., Norman, P. & Allen, T. (1982) Computer-assisted measurement and identification of honeybees (Hymenoptera: Apidae). *Annals of the Entomological Society of America*, 75, 591-594.

Dayrat, B. (2005) Towards integrative taxonomy**.** *Biological Journal of the Linnean Society,* 85**,** 407–415.

Deans A.R. & Kawada R. (2008) *Alobevania*, a new genus of neotropical ensign wasps (Hymenoptera: Evaniidae), with three new species: integrating taxonomy with the World Wide Web. *Zootaxa*, 1787, 28–44.

Dietrich, CH. & Pooley, CD. (1994) Automated identification of leafhoppers (Homoptera: Cicadellidae: *Draeculacephala* Ball). *Annals of the Entomological. Society of America*, 87, 412–423.

Drake, J. (2009) Robotic Automated Pest ID. In *CPHST 2008 Annual Report*. In prep. USDA-APHIS-PPQ-CPHST, Raleigh, North Carolina.

Eshera, MA & Fu, KS. (1986) An Image Understanding System Using Attributed Symbolic Representation and Inexact Graph-Matching. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 8(5), 604-618.

Evenhuis, N.L. (2007) Helping Solve the "Other" Taxonomic Impediment: Completing the Eight Steps to Total Enlightenment and Taxonomic Nirvana. *Zootaxa*, 1407, 3–12.

Fedor, P., Vaňhara, J., Havel, J., Malenovsky, I. & Spellerberg, I. (2009) Artificial intelligence in pest insect monitoring. *Systematic Entomology,* 34, 398-400.

Gaston KJ & O'Neill MA. (2004) Automated species identification: why not? *Philosophical Transactions of The Royal Society Of London Series B-Biological Sciences*, 359(1444), 655–667.

Gauld, I. D., O'Neill, M. A. & Gaston, K. J. (2000) Driving Miss Daisy: the performance of an automated insect identification system. In *Hymenoptera: evolution, biodiversity and biological control* (ed. A. D. Austin & M. Dowton), pp. 303–312. Collingwood, VIC: CSIRO.

Godfray, H.C.J. (2002a) Challenges for taxonomy. *Nature*, 417, 17–19.

Godfray, H.C.J. (2002b) Towards taxonomy's 'glorious revolution'. *Nature*, 420, 461.

Godfray, H.C.J. & Knapp, S. (2004) Introduction. [One contribution of 19 to a Theme Issue 'Taxonomy for the twenty-first century']. *Philosophical Transactions of the Royal Society of London, B* (2004), 359, 559–569.

Gonzalez, R.C. & Woods, R.E. (2007) *Digital Image Processing*. 3rd edition. Harlow: Pearson/Prentice Hall. xxii, 954 pp.

Hartley, C.J., Newcomb, R.D., Russell, R.J., Yong, C.G., Stevens, J.R., Yeates, D.K., La Salle, J. & Oakeshott, J.G. (2006) Amplification of DNA from preserved specimens shows blowflies were preadapted for the rapid evolution of insecticide resistance. *Proceedings of the National Academy of Science*, 103(23), 8757–8762.

Hebert P.D.N., Cywinska A., Ball S.L. & deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, 270, 313–321.

Hebert, P.D.N., & Gregory, T.R. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology*, 54, 852–859.

Houle, D., Mezey, J., Galpern, P. & Carter, A. (2003) Automated measurement of *Drosophila* wings. *BMC Evolutionary Biology*, 3:25, 1–13.

International Commission on Zoological Nomenclature (2008) Proposed amendment of the *International Code of Zoological Nomenclature* to expand and refine methods of publication. *Zootaxa*, 1908, 57–67.

Johnson, N.F., Masner, L., Musetti, L., van Noort, S., Rajmohana, K., Darling, D.C., Guidotti, A. & Polaszek, A. (2008) Revision of world species of the genus Heptascelio Kieffer (Hymenoptera: Platygastroidea, Platygastridae). *Zootaxa*, 1776, 1–51.

Jonker, R., Groben, R., Tarran, G., Medlin, L., Wilkins, M., Garcia, L., Zabala, L. & Boddy, L. (2000) Automated identification and characterisation of microbial populations using flow cytometry: the AIMS project. *Scientia Marina*, 64, 225–234.

Klinker, GJ., Shafer, SA. & Kanade, T. (1990) A Physical Approach to Color Image Understanding. *International Journal of Computer Vision*, 4(1), 7–38.

Liu, B., Maier, D. & Manner, R. (2005) An efficient and accurate method for 3D-point reconstruction from multiple views. *International Journal of Computer Vision*, 65, 175–188.

Lowe, DG. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91-110.

MacLeod, N. (ed) (2007) Automated Taxon Identification in Systematics: Theory, Approaches and Applications. Systematics Association Special Volume, 74. Boca Raton: Taylor & Francis.

May, R.M. (2004) Tomorrow's taxonomy: collecting new species in the field will remain the rate-limiting step. *Philosophical Transactions of the Royal Society of London, B*, (2004) 359, 733–734.

O'Neill, M.A. (2007) DAISY: A practical computer-based tool for semi-automated species identification.,In. MacLeod, N. (ed) Automated Taxon Identification in Systematics: Theory, Approaches and Applications. Systematics Association Special Volume, 74, pp. 101-114. Boca Raton: Taylor & Francis.

O'Neill, M. A., Gauld, I. D., Gaston, K. J. & Weeks, P. J. D. (2000) Daisy: an automated invertebrate identification system using holistic vision techniques. In *Proceedings of the Inaugural Meeting BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)* (ed. D. Chesmore, L. Yorke, P. Bridge & S. Gallagher), pp. 13–22. Egham: BioNET-INTERNATIONAL Technical Secretariat.

Pyle, R.L., Earle, J.L. & Greene, B.D. (2008) Five new species of the damselfish genus *Chromis* (Perciformes: Labroidei: Pomacentridae) from deep coral reefs in the tropical western Pacific. *Zootaxa*, 1671, 3–31.

Reed, TR. & Hans du Buf, JM. (1993) A Review of Recent Texture Segmentation And Feature-Extraction Techniques. *CVGIP: Image Understanding*, 57(3), 359–372.

Rohlf, FJ. & Bookstein, FL. (2003) Computing the Uniform Component of Shape Variation. *Systematic Biology*, 52, 66–69.

Russell, K.N, Do, M.T., Huff, J.C. & Platnick, N.I. (2007) Introducing SPIDA-Web: wavelets, neural networks and internet accessibility in an image-based automated identification system. In. MacLeod, N. (ed) Automated Taxon Identification in Systematics: Theory, Approaches and Applications. Systematics Association Special Volume, 74, pp. 131-152. Boca Raton: Taylor & Francis.

SOS Report (2009). State of Observed Species Report 2009. International Institute for Species Exploration, Arizona State University, in partnership with International Commission on Zoological Nomenclature, International Plant Names Index, Thomson Reuters and International Journal of Systematic and Evolutionary Microbiology. 10 pp.
http://species.asu.edu/files/IISE_SOS_2009.pdf

Stork, N. E. (1999) The magnitude of biodiversity and its decline. In J. Cracraft & F. Grifo (eds) *The Living Planet in Crisis: Biodiversity, Science and Policy,* pp 3–32. Columbia University Press, New York.

Suarez, A.V. & Tsutsui, N.D. (2004) The value of museum collections for research and society. *BioScience*, 54, 66–74.

Tautz D., Arctander P, Minelli A, Thomas RH, Vogler AP. (2003) A plea for DNA taxonomy. *Trends in Ecology and Evolution*, 18, 70–74.

Tofilski, A. (2004) DrawWing, a program for numerical description of insect wings. *Journal of Insect Science*, 17, 1–5.

Walters, T., Scher, J. & Drake, J. (2008) Identification Technology Program in Review. In *CPHST Laboratory Fort Collins: 2007 Annual Report*, pgs. 30-37. USDA-APHIS-PPQ-CPHST, Fort Collins, Colorado.

Wheeler Q.D., Raven, P.H. & Wilson, E.O. (2004) Taxonomy: impediment or expedient? *Science*, 303, 285.

Wheeler, Q.D. (2007) Invertebrate systematics or spineless taxonomy? *Zootaxa*, 1668, 11–18.

Wheeler, Q.D. (ed.) (2008) The New Taxonomy. The Systematics Association Special Volumes Series. 76. Boca Raton : CRC Press

Will, K.W., Mishler, B.D. & Wheeler, Q.D. (2005) The Perils of DNA Barcoding and the Need for Integrative Taxonomy. *Systematic Biology*, 54, 844–851.

Wilson, E.O. (1985) The biodiversity crisis: a challenge to science. *Issues in Science & Technology*, 20–29.

Wilson, E.O. (2004) Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London, B* (2004), 359, 739.

Winterton S.L. (2009) Revision of the stiletto fly genus *Neodialineura* Mann (Diptera: Therevidae): an empirical example of cybertaxonomy. *Zootaxa*, 2157, 1–33.

Zhang, Z.-Q. (2006) The making of a mega-journal in taxonomy. *Zootaxa*, 1385, 67–68.

Zhang, Z.-Q. (2008a) Zoological taxonomy at 250: showcasing species descriptions in the cyber era. *Zootaxa*, 1671, 1–2

Zhang, Z.-Q. (2008b) Contributing to the progress of descriptive taxonomy. *Zootaxa*, 1968, 65–68.