Bob Mesibov fra Tasmanien

TAXONOMY SHOULD BE DISTRIBUTED

The proposal is to distribute taxonomy in much the same way as the engineering of open-source software is distributed.

This is a very different path from the one usually put forward as a solution to the taxonomy crisis.

The customary call is to start more university graduates in careers as taxonomists. After appropriate undergraduate and graduate training, the budding taxonomists are expected to get jobs as embedded staff in museums, herbaria and universities, or perhaps (more vaguely) to join multidisciplinary, task-focussed teams. This strategy might be called More Career Taxonomists (MCT)

What's wrong with MCT? Two things.

First, it is totally inadequate as a strategy to vastly increase the world's taxonomic effort. The best it can hope to do, given realistic limits on what funding is available for training and hiring, is to slow the attrition of the taxonomic workforce. A highly optimistic view is that MCT could lead to a steady state, where the number of new career taxonomists finding jobs each year equals the number losing jobs or going into non-productive retirement.

Second, advocates of MCT see the fundamental unit of taxonomy as a person, i.e. a taxonomist. This one person might be a collector, nomenclaturist, morphologist, molecular geneticist, phylogeneticist, biogeographer, writer and illustrator - all at the same time. However, playing these diverse roles should not be confused with doing taxonomy. The fundamental unit of taxonomy is not a person, but a documentary record, e.g. a valid publication under a Code or a specimen identification in a museum collection database.

We do not need to increase the number of career taxonomists in order to increase the number of documentary records. The alternative to MCT is to tease out the component tasks of taxonomy and to set up distributed networks of people who specialise in those tasks. These networked contributors do not need to be taxonomists, nor for that matter university graduates. A network member could be a schoolboy in Belgium, a grandmother in California, a long-serving prisoner in Japan, an accountant in Uruguay.

The proposal is to divide the overall task of taxonomy - which is to discover and document the world's biota - into much smaller tasks, and to do these smaller tasks as well as possible. There is a parallel here with computer operating systems. MET is the Windows approach, where the OS is so tightly integrated as to be monolithic. The distributed approach is the one used in Unix-type systems: each component does one small job, but does it exceptionally well. Windows is expensive and requires an army of paid technicians to keep it afloat. Unix-type systems are free and are improved 24/7 by a vastly larger army of largely unpaid programmers.

The analogy can be taken one step further. The proposal is not to 'arm the ignorant' with tools to publish their own faulty taxonomic papers and throw taxonomy into chaos. As with sequential releases of the Linux kernel and many other successful open-source projects, the taxonomic effort can be organised hierarchically, with 'bug-checks' and data validation at several levels before the documentary record is released. A single expert compiling a monograph will not do this job as quickly and efficiently as a hierarchically organised team of 'citizen scientists' acting as task specialists and assistants.

What are the tasks of taxonomy? The traditional ones are easy to distinguish, and I list them below. However, it is not at all obvious that this is the division best suited to a distributed taxonomic network; there could well be better ones.

1. Backgrounding

a. Reviewing the literature

b. Reviewing existing collections

2. Working up

a. Fieldwork

b. Getting data from specimens

c. Data analysis (of characters, biogeography, etc)

d. Type management

3. Documentation

a. Naming under a Code

b. Managing synonymies

c. Diagnoses and descriptions

d. Illustration, key creation

e. Specimen curation and databasing

Under MCT, all or nearly all of these task classes would remain the responsibility of a single person. What has been done in recent years to streamline the work is to give that one person easier access to the raw materials for a particular task. Examples are the growth of digitised, online taxonomic literature and image catalogs of type specimens, and software (e.g. DELTA) for organising taxonomic data and generating diagnoses, descriptions and keys. Even in this approach we are well behind. It seems to me quite incredible that in 2008 there is no online resource for devising, validating and priority-checking Linnean binomens. Many taxonomists apparently still rely on advice from scholars of the Greek and Latin classics. (ZooBank was planning to offer a validator and priority-checker, but not a dictionary/thesaurus.)

  From: Bob Mesibov (mesi...@southcom.com.au)

  Date: 04/07/2008 11:44:31 PM

  List: edu.ku.nhm.mailman.taxacom

Hi, Gordon.

Sorry, I can't help with funding, and I'm not sure who can. A serious obstacle to getting a global sampling going is that expert fieldwork is competitive. There are only X people currently capable of doing it, and Y dollars to pay them.

It isn't yet obvious to me how sampling can be freed from this competitive arena. Maybe other TAXACOMers have examples? I keep my eyes **open** for 'piggyback' opportunities. For example, here in Australia the Royal Geographical Society of Queensland decided to get all sorts of folks to go the major lat/long intersections in the State, and document these 180-odd spots:

http://www.rgsq.gil.com.au/qldbydeg.htm

People, mainly non-expert locals, will be going to these spots and taking notes and photos. I contacted RGSQ to ask that the spot-visitors grab flora and fauna as well, and I offered to act as an intermediary with Queensland Museum for collection protocols and sample curation.

Nothing's happened yet, but even if sampling is incomplete and irregular, it would be a start in recruiting samplers from the world outside Expert-dom.

Maybe a better example of distributing taxonomic tasks is character/sequence analysis. Taxonomists prepare character matrices and sequence comparisons. They use these to generate classifications and/or phylogenies using a wide range of software programs. They do this themselves, and they publish the results as their own work.

For heaven's sake, why? The goal is to produce a working classification/phylogeny, a least-implausible tree given the inputs. Who cares whether Fred Bloggs and assistants did it, or an online server dedicated to the Bayesian analysis du jour, something like the validation parsers set up by the W3C to check Web-page code?

To truly free up this process, we need a motley group of thinkers to crystallise around building, testing and tinkering with the analytical software. These people will argue just as loudly as the cladists and pheneticists have in past years, with the difference that as part of a tree-improvement network they will only be competing for results and in-group recognition, like so many of today's tens of thousands of **open-source** programmers, and not for academic prestige and citations. These analysis enthusiasts will run and program the analytical software servers.

You, the taxonomist, generate the raw data. You feed it raw online to the tree-improvement network. You might get a message back saying 'Crap. Get more data.' You might get a tree back built with a method you don't like, but which is a less implausible tree than one built with a method you like. The results from your data might intrigue network members for technical reasons unrelated to biology, leading to new tweaks and new directions in inference. The raw data, the programs playing with them, and the results are all universally available online. The latter feed into an automatically updated (and expert-moderated) online Tree of Life.

You also publish your tree with the acknowledgement: 'Generated 11.05 GMT on 26 October 2009 by the Systematics Server'. Hours you would otherwise have spent learning about and tinkering with mindbogglingly complex, frequently updated inferential models can instead be spent generating more data - or discovering new taxa.

For years I've been told that statistical phylogenetic inference is non-subjective, reproducible, robust, etc. Sounds like machine work to me. Why waste your time doing it? Why not give it to distributed amateur specialists who think about nothing else in their spare time, and who couldn't give a fig whether A is more closely related to B or C, or whether a clade is monophyletic or paraphyletic?

I hear flame-thrower safeties being thumbed off...

-- Dr Robert **Mesibov** Honorary Research Associate, Queen Victoria Museum and Art Gallery and School of Zoology, University of Tasmania Contact: PO Box 101, Penguin, Tasmania, Australia 7316 (03) 64371195; 61 3 64371195 http://www.qvmag.tas.gov.au/mesibov.html