Opinion

Special Issue: Ecological and evolutionary informatics

Ramping up biodiversity discovery via online quantum contributions

David R. Maddison[1], Robert Guralnick[2], Andrew Hill[3], Anna-Louise Reysenbach[4] and Lucinda A. McDade[5]

[1] Department of Zoology, 3029 Cordley Hall, Oregon State University, Corvallis, OR 97331, USA
[2] University of Colorado Museum of Natural History and Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, Boulder, CO 80309, USA
[3] Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, Boulder, CO 80309, USA
[4] Biology Department, Portland State University, Portland, OR 97201, USA
[5] Rancho Santa Ana Botanic Garden and Claremont Graduate University, 1500 N. College Ave., Claremont, CA 91711, USA

The pace of species discovery and documentation remains too slow on a human-altered planet in the midst of a massive extinction event. Increasing this pace requires altering conventional workflows. In this review, we propose that systematics needs to shift to a model of quantum contributions whereby species hypotheses are published as they are formulated and data as they are collected in web-based repositories and content-management systems. If our recommendation is followed, many species will make their first appearance on the Internet as candidate new species before documentation is complete. Acknowledging the changes that we describe may be controversial, we discuss problems that may be encountered along with possible solutions.

Too many species, too much extinction, too little time

The discovery and documentation of biodiversity on Earth is proceeding at an inadequate pace, especially in the most diverse groups. Approximately two million species have been documented so far and an order of magnitude more may remain to be discovered (http://www.catalogueoflife.org/annual-checklist/2009) [1–3]. An estimated 6200–18 000 species of eukaryotes are described per year [3,4]. These estimates yield the frightening prospect of a millennium of basic discovery work at the present pace. In addition, anthropogenic impacts are dramatically altering the biota of the Earth, with extinction rates now as high as 27 000 known species per annum [5,6]. Biologists need to embrace biodiversity discovery and documentation with increased urgency, commitment and innovation: if the *status quo* is maintained, it will take too long. Many improvements can be made to the way in which biodiversity is discovered and documented that will increase the pace [7]. Here, we focus on what we view as a vital change: altering how and when discoveries are shared. We believe that species discovery will proceed most rapidly if data and species hypotheses are published as they are generated in Web-based data repositories. We articulate the advantages of this approach below and in Box 1.

Conventional paths to species discovery and documentation are time-consuming and risky

Any method of discovering species will be time-consuming when a taxonomic group is species-rich, occurs in logistically challenging areas, or requires special techniques to collect or acquire the necessary data. The traditional approach of producing large revisionary works slows the pace further and puts that work at risk of loss for the following reasons: (i) The traditional process of species discovery and documentation is lengthy and few tangible products are published along the way. Projects often take multiple years and, thus, are vulnerable to events that impede their completion. (ii) Until recently, much of this research has occurred in systematists' minds and on paper. Lamentably, both minds and paper are subject to damage and demise. Knowledge is therefore vulnerable as careers inevitably end. Expertise gained through a lifetime may be lost before it is put down in a publicly available format. (iii) Taxonomic research has often been undertaken by individuals working in isolation. Solo research has advantages, accommodating the style of some systematists, but collaborative efforts can achieve synergies that speed and enhance research products. Similarly, working in isolation leads to missed opportunities to build the contributions and knowledge of others into research products.

Smaller, more frequent contributions will speed species discovery and documentation

Knowledge of biodiversity increases through dynamic interplay between data and hypotheses. For macro-organisms, it begins with collecting and acquiring data from specimens; for microbes, the first data may come from DNA sequences. In either case, data are obtained that enable initial assessments of similarity compared to known relatives. At some point, a researcher obtains the first hints that data may represent an unrecognized species. In the next phase, the hypothesis is tested with more data, further refined, and so on. At a point along this path, the investigator becomes confident that the data represent a distinct species, after which the species receives a name. More data might be gathered after that point, and hypotheses about species

Corresponding author: Guralnick, R. (Robert.Guralnick@colorado.edu)

 0169-5347/$ – see front matter doi:10.1016/j.tree.2011.10.010 Trends in Ecology and Evolution, February 2012, Vol. 27, No. 2

Box 1. Benefits of publishing systematic research through quantum contributions on the Internet

- Making knowledge of new species available more quickly before they are named or fully documented in revisionary works.
- Placing new data more rapidly in the context of existing data, and allowing them to be explored using synthetic tools linked to Internet repositories.
- Enabling collaborative work in which the complementary resources and talents of individuals can more quickly establish the existence of new species.
- Reducing the loss of systematic knowledge that now occurs owing to the demise of systematists and their personal digital or paper records.

boundaries further refined. Traditionally, publication is at the end of this process.

We argue that knowledge about biodiversity will increase most rapidly if data and hypotheses are made available throughout the process, as quantum contributions. A quantum contribution in systematics might be a photograph of a specimen; a new specimen record, including geographic information; a DNA sequence; a description of a diagnostic feature; or a hypothesis that a distinct species exists. Venues for publication of quantum contributions are Internet-based data management systems. Data providing evidence of new species will need to be as high quality, and as rigorously gathered and presented, as ever. Species hypotheses will need to be carefully advanced and tested. It is our contention that the 'publish-as-you-go' model we espouse will often lead to species hypotheses that are supported by more extensive data than is typical, and will have been vetted more efficiently at all steps in the process by the community.

One key advantage of quantum contributions is that they are easier for systematists to produce than are full revisionary publications and, thus, knowledge about biodiversity will enter the public sphere without long delays. Many systematists know of dozens of undescribed species that they plan to describe someday. If easy, comprehensive systems for quantum contributions are created, then spending a few minutes posting pictures or DNA sequences will make a positive and significant contribution to biodiversity knowledge, sometimes years before the systematist completes a formal species description.

We are also convinced that the quantum contribution model is essential for safe-guarding products of systematic research. Doing so protects data from loss owing to any number of circumstances, from demise of one's personal computer to demise of oneself. Projects in progress are far safer from total loss before completion if they are web-based than if they are only in one's desktop computer, notebooks, filing cabinets and mind.

We note that scientists who work on microbial diversity provide a proof of concept of this approach. This community focuses on sequencing and annotating organismal genes and genomes, and has championed and embraced the release of data before formal publication [8,9]. Rapid, open access to these data has allowed the attendant repositories (e.g. GenBank) to increase rapidly in size and complexity. Such repositories have spawned tools for managing, visualizing and analyzing sequence data, and this reciprocal development has greatly expanded knowledge of microbial diversity (Box 2). Similar tools geared for systematists working on other groups of organism are becoming available (e.g. the Barcode of Life Database; Box 2). Although current tools have limitations (Box 2), we argue that the rapid release of data will be similarly beneficial to documenting the 'genome' of Earth; that is, the organismal diversity of the world.

In the fields of astronomy and physics, the Sloan Digital Sky Survey (SDSS) [10] and ArXiv projects represent two remarkably successful experiments in providing immediate, freely available resources. These transformational successes, which have allowed new kinds of science (e.g. citizen science utilizing SDSS data [11]), are often discussed as models, but remain unreplicated in biodiversity sciences.

Quantum publication on the Internet

We agree that the future of publication of taxonomic results lies in the Internet, as cogently argued by Godfray *et al.* [12]. We also agree with Mietchen *et al.* [13] that Internet resources (e.g. wikis) can improve works after species have been formally named. We are promoting a more pervasive change: early Internet publication of quantum contributions about undiscovered or incompletely known organisms (and not just finished taxonomic works or modifications of these) has the potential to speed biodiversity discovery and documentation as it transforms our field.

Some digital tools already available can be components of a workflow via quantum contributions (Box 2). These tools provide ways to compile biodiversity information collaboratively and efficiently, while also allowing experts to share knowledge via annotations (e.g. expert opinions). Such digital environments will support the 'publish-as-you-go' research approach that we argue is vital to hastening biodiversity discovery and documentation. With further development and better, more open linkages across platforms, web-based tools such as these will allow individuals and communities to examine new data in context, quickly determine the current state of knowledge and prioritize precious resources to hasten species documentation.

Collaboration hastens discovery

The collaborations enabled by quantum contributions will speed species discovery and documentation. When one systematist posts an image of an odd specimen, another might recognize it as similar to something s/he has seen along a nearby river and then collect specimens for DNA sequencing and share results via GenBank. A third systematist might note similar specimens from elsewhere and publish distribution records. At that point, these systematists might communicate and decide that there is enough evidence to create an informal name for the proposed species. The workflow that we advocate will also facilitate contributions to the process by non-systematists and amateurs. With ready access to information about a new species, a field ecologist might contribute the observation that organisms of this species have an ecological association with another species. A paleontologist might compare

Box 2. Current informatics endeavors for biodiversity compilation: existing pieces of the quantum contributions puzzle

Social networking platforms for sharing biodiversity knowledge and resources for coordinating new data into existing phylogenetic contexts are rapidly being developed. Missing is the ability to integrate these different kinds of platform in ways that maximally support discovery within and across platforms, and that seamlessly link taxonomy, phylogenetic trees and the data objects that adorn the branches of these trees, and that have data-entry, visualization and presentation tools elegant enough to attract most systematists. Here, we highlight some current efforts that individually provide great utility.

Tools for collaboratively generating biodiversity content

Scratchpads
Scratchpads (http://scratchpads.eu/) provide a social networking platform to create, share and manage taxonomic data online. Especially valuable is automatic association of data uploaded to Scratchpads with Encyclopedia of Life (EOL) taxonomies. This is powerful for providing summary content of known taxa and beginning to associate existing, but not new, data into taxonomic frameworks.

Life Desks
Life Desks (http://www.lifedesks.org/) are similar to Scratchpads and seamlessly integrate with EOL taxon pages that include aggregated content from across the web. LifeDesk and EOL taxon pages include undescribed species and thus provide a mechanism to locate provisional taxa.

CATE
CATE (Creating a Taxonomic e-Science, http://www.cate-project.org/) is focused on developing unitary taxonomies overseen by the community, along with development of modern, web-based treatments of taxa.

Wikispecies
Wikispecies (http://species.wikimedia.org/wiki/Main_Page) is a free species directory, open to creation and editing by the public. It does not integrate with other databases, nor does it include automatic data-quality annotations.

Tools for phylogenetic ordination and taxonomic identification

Ribosomal Database Project
The Ribosomal Database Project (http://rdp.cme.msu.edu/) includes classifier and tree-building tools that allow new sequences to be ordinated in context with existing sequences, but lacks community workbenches to further utilize these results.

BOLD
BOLD (http://www.boldsystems.org), the Barcode of Life Database, provides identification of unknown sequences and placement within trees but lacks functionality to utilize this output fully.

Metagenomics databases and tools
Metagenomics databases and tools, such as Camera, MG-RAST and IMG/M, are proliferating and provide both phylogenetic analysis tools and binning methods to make first-cut taxonomic matches. However, these are not specifically geared toward doing biodiversity discovery and documentation.

Workflows for speeding documentation and publication processes
Novel approaches for hastening movement from provisional assessment of new units of biodiversity to fully documented units are also being developed. In particular, automated workflows, such as those documented by Blagoderov *et al.* [25], extend Scratchpads projects with name registration in ZooBank and publication in ZooKeys.

his/her unidentified fossils to an image posted online and realize that s/he has collected a related species. An amateur who grows or raises similar species might contribute images, a process that is already in place for mushrooms (http://mushroomobserver.org/). By contrast, a lone systematist might take years to arrange a field trip and might not be aware that another systematist has collected specimens, might not have the resources to visit that collection to study the specimens, or might not have access to images of fresh material supplied by the amateur. Thus, the 'publish-as-you-go' approach will increase the data supporting species hypotheses while decreasing the time between the first hints of a new species and public recognition of such.

An ever-growing, global, phylogenetic revision

Data that are contributed to biodiversity databases must be presented in the context of current knowledge. One mechanism would be to attach data that may denote new species to an Internet-based depiction of the tree of life that summarizes existing knowledge of organismal diversity and phylogeny. Standardized and valid taxon names provided through various services would serve as a backbone classification for this organized knowledge. Newly obtained data would be attached to this tree along with existing evidence, be it DNA sequences, image of plumage, or video of species-specific behavior. Once evidence suggests a new species, that species would receive an informal name and its own node on the tree, at which all data about the species would be accumulated. The milestone of creating the placeholder node for a species in the cyber tree, along with its informal name, would be followed by creation of a formal taxonomic name and complete documentation. The result would be a continuously updated revision of all life. This novel workflow is summarized in Figure 1.

Taxonomic milestones and quantum contributions: species hypotheses, informal and formal names

A unique challenge for systematics in a quantum contribution-based knowledge model is accommodation of classifications and taxonomic names. These are not only associated closely with scientific hypotheses about phylogeny and species boundaries that systematists infer, but must also satisfy rules of nomenclature that are designed to facilitate communication.

Traditionally, species are not mentioned in literature until they have formal names, and species are usually named only after they are well documented. Recent advances in publishing workflows (Box 2) hasten the final steps in the current documentation process. In the culture of publishing quantum contributions, however, the first data would often be published before any researcher has enough confidence to create a formal name. Data may even be available before the researcher suspects they represent a new species. Either way, information about species would be publicly visible while the new species has at most an informal name.

Three taxonomic elements will thus be created during the process: (i) a species hypothesis or concept: a claim that a species exists that consists of a defined set of specimens and related data; (ii) an informal name or taxon label

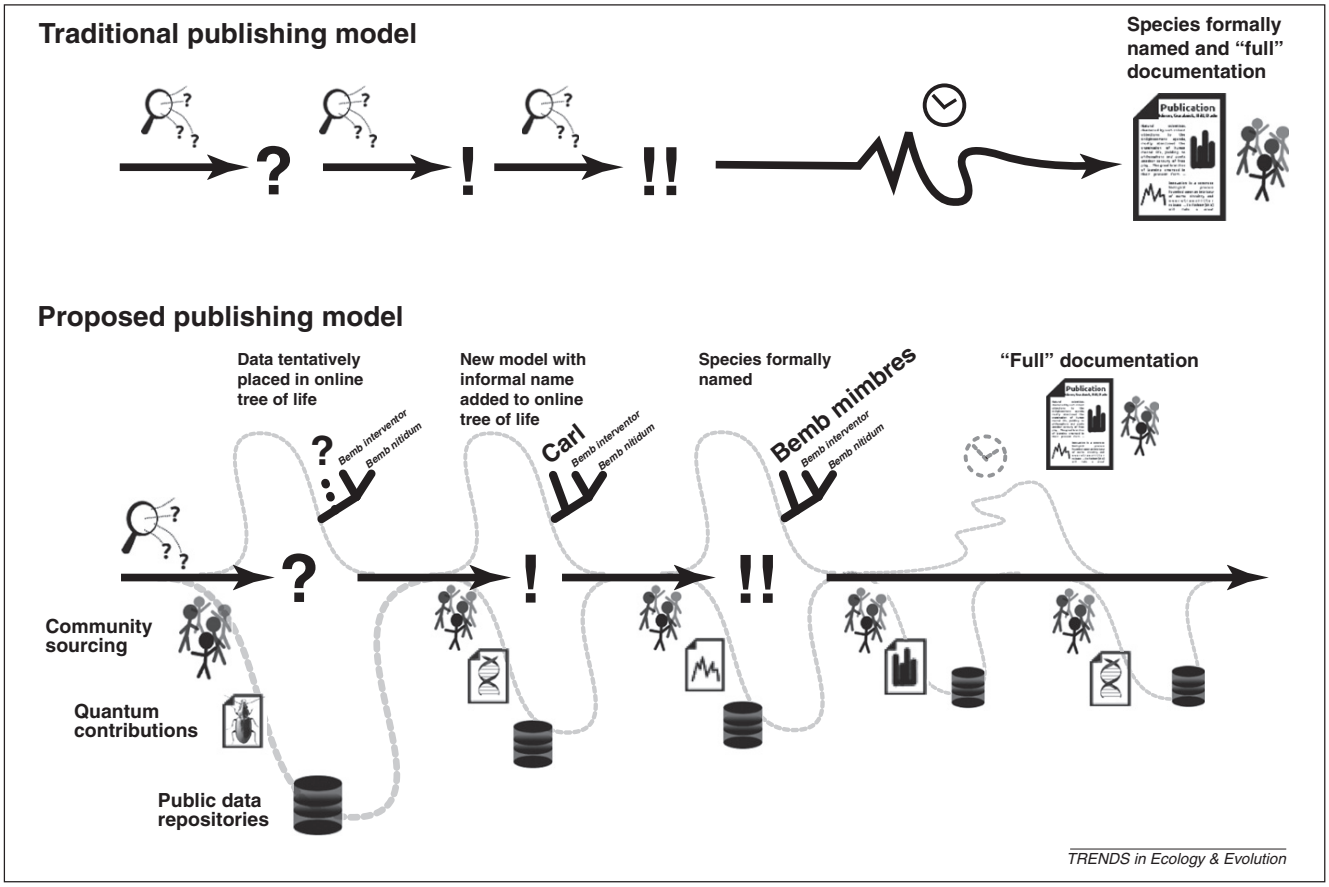

Figure 1. Traditional publishing model compared to the proposed publishing model. In the traditional model, publication, including formal naming of a species, occurs after a long period of solitary research, with the results available to the community at large only at the end. The workflow we advocate for species discovery and documentation starts, as in the past, with generating initial data suggesting a possible new unit of biodiversity, shown as a '?' above. The next steps are quite different, however, including publishing all initial and new data into web-based repositories, thus making these data rapidly available and discoverable to other researchers for their additions. In the proposed publishing model, anyone could be part of the community contributing data, with annotations to indicate data quality; a slightly less-radical model would entail restricting contributions to a select few. These data are linked onto a growing tree of life, beginning with tentative placement on the tree, then creation of a new, informally named node on the online tree of life when researchers have enough evidence to propose that it represents a new species ('!'). Confidence in the distinctiveness of the new species reaches a point ('!!') where it is formally named. This can later be followed by a publication documenting the species in more detail.

associated with vouchers [14]; and (iii) a formal taxonomic name satisfying Code-governed nomenclatorial rules. In web-based repositories, several pieces of metadata would be attached to each element, including a globally unique identifier (GUID) or code to specify each element uniquely. Schindel and Miller [14] have argued for formalizing informal names as taxon labels to ensure that they are unique and stable; an equivalent solution would be GUIDs as components of both informal and formal names. Such identifiers and semantic descriptions of data provide means to track names and enable linkages between them and other data. This vision, that creation of a species hypothesis, informal names and formal names are three distinct milestones in documenting a species, is in contrast to standard taxonomic practice of conferring formal names only simultaneously with publication of species hypotheses (Figure 1).

Data publication before and after formal naming

For some communities, publishing data on species before they are formally named is already the cultural norm. In black flies, for example, new species are often discovered via novel polytene banding patterns and given informal names before they receive formal names (e.g. *Simulium appalachiense* Adler, Currie and Wood [15] was formally named 42 years after it was described as *Simulium tuberosum* 'CDE sibling' [16]). More recently and generally, new species are being discovered and DNA sequences published before they are described formally (e.g. in fungi [17], beetles [18], meiobenthic organisms [19], Bacteria [20] and Archaea [21,22]).

Page coined the term 'dark taxa' for taxa represented by DNA sequences in GenBank that do not yet have formal species names (http://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html). Some represent described species that have not been so identified, others represent new species recognized by researchers but not yet formally named, and the remainder await further research to determine their status. These dark taxa in GenBank are on varied but early parts of the pathway of a faster, invigorated taxonomic workflow; their eventual acquisition of formal names will help complete that path.

Once a formal taxonomic name is applied, should documentation continue until it is as complete as in traditional species descriptions? This will be determined by community

needs. Some communities might decide that, as photographs of type specimens, DNA sequences and other data become more prevalent, gathering some traditional data has lower priority than moving on to discover additional species. Thus, the approach we outline would also increase the pace of discovery by reducing less-critical documentation.

Conclusions and caveats

We have described a process to speed and safeguard systematic work that emphasizes publishing data and hypotheses as they are acquired. Not all systematists will adopt the fully open approach proposed. The optimal approach may differ from clade to clade, as a function of the nature of the community that works on that clade, the popularity of the clade and the amount of available data. Some communities may initially feel more comfortable only sharing data among known, trusted members. We believe, however, that progress will be most rapid in those clades for which the community embraces a publish-as-you-go model of quantum contributions.

We reiterate that traditional approaches to biodiversity discovery and documentation are not rapid enough and leave critical knowledge vulnerable to loss. However, we recognize that what we espouse is not problem-free and we briefly discuss some issues below.

A key current limitation is a lack of tools for contributing, synthesizing and presenting digital data that are complete and user-friendly enough to be compelling to most systematists. Digital resources will not dramatically speed biodiversity discovery and documentation unless their benefits, in terms of knowledge and credit gained, outweigh the costs of contributing (the topic of professional 'credit' for research activities of systematists is addressed elsewhere [23]). Furthermore, once data are entered, there is at present no 'full-service' platform that synthesizes them into products that systematists find fully satisfactory, such as keys and identification tools, elegant species pages with images and distribution maps, phylogenetic trees and pages for higher-level taxa. User-friendly, globally linked, dynamically updating and interoperable resources will need to be completed to provide the foundation of the fast-paced systematics culture that we imagine.

We foresee two other main concerns. First, given the amount of unverified 'information' available over the Internet and fears that crowd sourcing yields material reflective of the lowest common denominator, some will be skeptical of adding content in any but a tightly controlled fashion. We are confident that by providing the community with rating and feedback tools (e.g. Stack Overflow: http://stackoverflow.com/ or PageRank [24]), and the ability to monitor deletions and revisions of data (e.g. Git commits: http://git-scm.com/, or Wiki histories), high-quality additions and annotations will rise to the top. Additionally, users will accrue status and reputation via consistent high-quality contributions, promoting or enforcing minimum standards for quality and providing oversight to minimize low-quality additions, much as peer review does at present.

Second, problems involving when and by whom species are named pose serious challenges. As data accumulate, there could be competition to create names. One researcher might post information about a new species, intending to name it eventually. A second researcher might see the online data and quickly plant the flag of a new species name on these data. The current rules of nomenclature would confer authorship on the second, usurping researcher. Fear of loss of nomenclatorial authorship might be one of the greatest deterrents to the cultural change that we espouse. We believe that such concerns will be allayed by new systems for tracking credit for data, the development of rules such as those used in the genomics community to govern early release of data [9], and perhaps changes in nomenclatorial rules themselves. We also note that the current system is not immune to such abuses; the more open, transparent system advocated may more effectively expose usurpers, allowing for faster community response.

Finally, in contending that members of the taxonomic community must change how they work, we by no means diminish the research achievements of systematists nor suggest that insights acquired over years of study can be replaced with technology. We recognize that the expertise necessary to discover and document biodiversity accrues over long periods of time and that this knowledge generated is not readily replaced. Precisely because such knowledge is so valuable, we argue that it is vital to commit to better ways of safe-guarding, sharing and publishing research results as they are generated. In so doing, it will be possible to more rapidly advance understanding of biodiversity on Earth.

Acknowledgments

Many thanks to the National Science Foundation and its program officers, especially Maureen Kearney, Scott Snyder, Rafael De Sa, Elizabeth A. Kellogg and Timothy Collins, for sponsoring and attending the workshops on Systematics and the Future of Biodiversity that inspired this contribution. We would especially like to thank workshop leaders and attendees, notably Patrick Herendeen, for their efforts, discussions and thoughts. These workshops were supported by NSF grant DEB 0935231 to Patrick Herendeen, Lucinda A. McDade and Petra Sierwald. Thanks, too, to John Lundberg, Wayne Maddison, Scott Miller, Roderic Page, David Schindel and three anonymous reviewers for providing thoughtful comments on the manuscript.

References

1 Chapman, A.D. (2009) *Numbers of Living Species in Australia and the World*, (2nd edn), Australian Biological Resources Study
2 Whitman, W.B. (2009) The modern concept of the prokaryote. *J. Bacteriol.* 181, 2000–2009
3 Mora, C. (2011) How many species are there on earth and in the ocean? *PLoS Biol.* 9, e1001127
4 International Institute for Species Exploration (2010) *State of Observed Species 2010*, International Institute for Species Exploration
5 Wilson, E.O. (1992) *Diversity of Life*, Harvard University Press
6 Sax, D.F. and Gains, S.D. (2008) Species invasions and extinction: the future of native biodiversity on islands. *Proc. Natl. Acad. Sci. U.S.A.* 105 (Suppl. 1), 11490–11497
7 Berger, J.K. (2005) Mission possible: ALL Species Foundation and the call for discovery. *Proc. Calif. Acad. Sci.* 65 (Suppl. 1), 114–118
8 Bentley, D.R. (1996) Genomic sequence information should be released immediately and freely in the public domain. *Science* 274, 533–534
9 The Wellcome Trust (2003) *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*, The Wellcome Trust
10 York, D.G. *et al.* (2000) The Sloan Digital Sky Survey: Technical summary. *Astronomical J.* 120, 1579–1587
11 Raddick, M.J. and Szalay, A.S. (2010) The universe online. *Science* 329, 1028–1029
12 Godfray, H.C.J. *et al.* (2007) The Web and the structure of taxonomy. *Syst. Biol.* 56, 943–955
13 Mietchen, D. *et al.* (2011) Wikis in scholarly publishing. *Inf. Serv. Use* 31, 53–59

14 Schindel, D.E. and Miller, S.E. (2010) Provisional nomenclature: the on-ramp to taxonomic names. In *Systema Naturae 250 – The Linnaean Ark* (Polaszek, A., ed.), pp. 109–115, CRC Press
15 Adler, P.H. *et al.* (2004) *The Black Flies (Simuliidae) of North America*, Comstock Publishing
16 Landau, R. (1962) Four forms of *Simulium tuberosum* (Lundstr.) in southern Ontario: a salivary gland chromosome study. *Can. J. Zool.* 40, 921–939
17 Arnold, A.E. *et al.* (2000) Are tropical fungal endophytes hyperdiverse? *Ecol. Lett.* 3, 267–274
18 Monaghan, M.T. *et al.* (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philos. Trans. R. Soc. B* 360, 1925–1933
19 Markmann, M. and Tautz, D. (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philos. Trans. R. Soc. B* 360, 1917–1924
20 Rappé, M.S. *et al.* (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418, 630–633
21 Könneke, M. *et al.* (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437, 543–546
22 Reysenbach, A-L. *et al.* (2006) Isolation of a ubiquitous obligate thermoacidophilic archaeon from deep-sea hydrothermal vents. *Nature* 442, 444–447
23 McDade, L.A. *et al.* (2011) Biology needs a modern assessment system for professional productivity. *Bioscience* 61, 619–625
24 Page, L. *et al.* (1999) *The PageRank Citation Ranking: Bringing Order to the Web, Technical Report*, Stanford InfoLab.
25 Blagoderov, V. *et al.* (2010) Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. *ZooKeys* 50, 17–28