

E-publishing descriptive taxonomy: the convergence of taxonomic journals and databases

Vincent S. Smith

Department of Entomology, The Natural History Museum, Cromwell Road, London, SW7 5BD, United Kingdom. (vince@vsmith.info)

Abstract

Descriptive morphological taxonomy remains a slow, labour-intensive undertaking. The number of specialists willing and able to focus their research time on any single taxon is limited and for an increasing number of taxa this expertise is no longer available. In order to pursue large-scale species discovery and description, the taxonomic community needs tools that accelerate the taxonomic process, maximise the efficiency of the time researchers invest, and embrace methods of publication that enable the recombination and reconfiguration of taxonomic data. Critically the tools must support the needs of a wider community of scientists beyond the taxonomic community. A major bottleneck in the taxonomic workflow is the time consumed by the interplay between the taxonomist and the publisher in preparing taxonomic data and going to print. Breaking this bottleneck requires seamless integration between compilation of the descriptive taxonomic data and the publication upon which the data are based. In this chapter I describe some of the major technical and social barriers that the taxonomic community has to address if we are to make this transition. I go on to describe selected projects that are in the early stages of overcoming these barriers to make this transition a reality. These systems are challenging the traditional concept of publication through the construction of Virtual Research Environments that enable the simultaneous compilation and publication of taxonomic data on the Web. Coupled with further changes in publishing technologies and social practices, I predict a gradual convergence of taxonomic journals and databases as unified entities.

Introduction

No one perceives a database entry of, say, a specimen in a museum collection or a DNA sequence, as being as valuable as the scientific paper that describes it. Accordingly researchers are not (usually) awarded promotion based on the number of deposits made to a biological database or the annotations made to database record. Rather, current measures of peer recognition focus on the number of papers published, the citations these papers attract, and their resulting “impact” as determined through various metrics. Yet ironically, to the consumer at least, the database entry may be more valuable than a paper (Bourne, 2005). Databases containing the products of the taxonomic community (e.g., taxonomic names and concepts, specimen geospatial and temporal data, collection records, images, phenotypic and genomic character states and descriptions), can be more easily transmitted, reused and repurposed than much of the knowledge locked up within traditional descriptive papers. Indeed, the audience for these databases is likely to be far higher given the specialist nature and limited availability of most traditional taxonomic

works. This assertion is borne out by download statistics from potentially analogous biological databases such as Genbank and the Protein Data bank, which have shown substantial, if not exponential growth over the past decade (Pruitt et al., 2009).

Comparable databases for taxonomy on the scale of resources like Genbank do not currently exist. Instead descriptive taxonomy continues to be disseminated in dense treaties that are usually only physically and intellectually accessible to specialists that have material access to these works, and the knowledge required to interpret their content. In consequence, most taxonomy is effectively withheld from use for a wide range of scientific applications, even to the point that it often cannot be readily incorporated into further taxonomic study. Why should this be so? Why should the taxonomic community persist in publishing descriptive taxonomic data in ways that limit its use to all but a handful of specialists? In this chapter I explore this question and highlight some of the technical and social developments that are taking place within the research and scholarly publishing industry, which are bringing change to the practice and dissemination of descriptive taxonomy.

Dissemination of results is one of the critical steps in the practice of science, and taxonomists, like other scientists, do this in many forms, through conference presentations, invited lectures and most importantly through publication. Publication is simply the act of making something publicly known through the preparation and distribution of multiple identical copies. In a scholarly context, however, scientific publishing takes on a special meaning. In 1665 Henry Oldenburg established the principles of scientific publishing as founding editor of *Philosophical Transactions of the Royal Society*. This was the first English language scholarly journal and was set up to disseminate, register, certificate and archive content that we would today recognise as science (Hall, 2002). These tenets of scholarly publishing are still relevant today despite the fact that the World Wide Web has supplanted print on paper as the primary medium for disseminating scholarly works. Online versions of publications are typically presented either as HTML Web pages or as static PDF documents, with the Internet used primarily as a convenient distribution medium for the text. As the electronic embodiment of the static printed page, the PDF document is directly comparable to the first scholarly articles published by the Royal Society in 1665, but it is antithetical to the spirit of the Web, which can support the constant updating and improvement of the published information on a continual basis.

Compared to what many database developers have achieved in terms of data integration, comprehension through novel visualisation techniques, and real time collaboration of authors, most modern publishers have failed to embrace alternative forms of publishing afforded by new, and increasingly Web-based technologies (Brown, 2008). An often cited reason for this is that changing and updating content conflicts with one of the original tenets of publishing as a means of producing multiple, identical copies. It also makes peer review difficult, since this usually relies on small communities of experts to review changes, and this cannot usually be delivered in real time. Yet, neither of these challenges is insurmountable. Snapshot versions of dynamic documents can be created to facilitate citation, while unvetted or edited content can be flagged for subsequent review.

Tools like MediaWiki (2007) have for the most part solved these technical challenges many years ago. Perhaps other more commercial interests drive mainstream publishers to preserve the status quo. Nevertheless, within the context of descriptive taxonomy there are genuine issues that need to be addressed if new forms of publishing are to be considered.

In the following sections I highlight what I consider to be the major thematic challenges for the taxonomic community if we are to embrace new forms of publishing, and then examine the first tentative steps made by taxonomists to tackle these problems through new IT infrastructures. Many of these issues are relevant to the wider scientific community, and are not limited to the problems of taxonomic publishing.

Challenges for e-publishing taxonomy

These can be broadly divided into technical issues associated the development of technologies and interfaces that support the taxonomic process, and social issues that require consensus solutions amongst the taxonomic and wider community. Perhaps not surprisingly the social issues are the dominant features on this list and the technical issues from a computer science perspective are not especially complex, with the notable challenge of data integration.

Technical issues

Supporting the taxonomic workflow

This refers to the services and tools that capture the lifecycle of taxonomic research from inception to publication. Taxonomic research is built from two resources: samples of the organisms involved (either specimens or observations) and the legacy of taxonomic investigation found in the published literature. Supporting the taxonomic workflow requires repurposing these data in a digital environment that supports the scientific process of taxonomy. The core of this process is encapsulated in a feedback loop between hypotheses of characters and taxa (Johnson, 2010). Taxonomists group specimens into taxa on the basis of empirical observations about the characters they exhibit. The resulting taxonomic concepts and the degree to which specimens are congruent with them may suggest new characters and character states, leading to modifications in the proposed taxon concept. The process is iterative and open ended, until a stable equilibrium is reached, at which point the taxon concept is fixed in the mind of the author and data are published as part of a taxonomic revision. This revision may incorporate a summary of published literature, checklists, and identification keys, in addition to the formalised descriptions (diagnoses) of taxa and supporting data (e.g. images, distribution maps, materials examined).

Data supporting the taxonomic workflow are replete with structure and organisation, making them eminently suited to automated process that might dynamically compile revisionary publications in real time. Indeed many taxonomists draw on databases in the compilation of these publications, so the fact that this structure is lost on publication is a major lost opportunity. Since all of the data used to compile taxonomic revisions can be stored within a database (even natural language sections like the abstract) it is possible to

retain all the data structure and provide dynamic views of these datasets. This concept is not new. The DELTA (DEscription Language for Taxonomy) data format developed between 1971-2000 and has supported the automated compilation of taxonomic keys and revisionary publications for some time (Dallwitz, 1980). This has been used in the production of many hundreds of descriptive taxonomic publications. However, this is a one-way process geared toward print based publication and is relatively complex to use, only attracting a small fraction of taxonomic community that have the computing skills required to make use of the software (Walter and Winterton, 2007). Despite these problems the relative success of DELTA illustrates that the primary challenge for software developers supporting taxonomists is not in the development of the data architectures (databases and standards) storing the taxonomic data. Rather it is in the development of an integrated experience for the taxonomist supporting the entire taxonomic workflow from project inception to publication. Such systems need to be integrated with services that, where possible, supply the underlying taxonomic data. This enables taxonomists to reuse and repurpose information that has been captured elsewhere, and publish their new digital data back to these hubs to benefit others. Data infrastructures like the Global Biodiversity Information Facility (GBIF, 2010) for specimen observational data, and Biodiversity Heritage Library (BHL, 2010) for taxonomic literature are central to this effort, despite the fact that they are at present, woefully incomplete (see *Digitisation*).

Data integration

Integrating diverse sources of digital information is a major technical challenge for the taxonomic community (Page, 2008). Not only are we faced with numerous, disparate data providers, each with their own specific user communities, but also the information we are interested in is extremely diverse. For example, the Taxonomic Databases Working Group (TDWG) currently lists 654 different database projects (TDWG-Statistics, 2010). Similarly currently GBIF lists 10,556 datasets from 317 different data publishers (GBIF-Statistics, 2010). Combine these with mainstream bioinformatics databases like GenBank and PubMed, plus the taxonomic literature digitised by the BHL (currently 28 million pages), and the magnitude of the challenge becomes readily apparent. Efforts to integrate these data are central to maximising the efficiency of the taxonomic publishing.

Initial efforts at data integration have taken two distinct paths; one focusing on shared identifiers (principally taxonomic names) to link data together, the other focusing on shared use of data standards. Of course there is some overlap between these approaches, but they have come to represent different philosophies about the future of biodiversity informatics. Use of shared identifiers to determine whether two items of data refer to the same entity allow resources to be linked or mined for information without being specifically structured to support this activity. In contrast shared uses of data standards require detailed agreement amongst the data providers about what data are to be integrated, how they are structured, and the protocols by which they are shared. When implemented properly, data standards can support a very high level of integration. For example the Access to Biological Collections Data schema (ABCD, 2010) is a comprehensive standard supporting the integration of data about specimens and observations for about 700 data elements. However, the complexity of ABCD means that

simpler and less granular standards (e.g. the Darwin Core 2010, with about 40 elements) receive in much wider use (Constable et al., 2010).

Taxonomic names as identifiers have received a great deal of attention, notably underpinning efforts in the Biodiversity Heritage Library to aggregate information within digitised taxonomic literature (Rinaldo, 2009). However, taxon names have a serious limitation as identifiers since they are neither stable nor unique. The name *Morus* for example, might equally refer to the seabird genus of the Gannet, or the plant genus for the Mulberry tree. A 250-year legacy of taxonomic research means that taxonomy is replete with such examples, and consequently taxon names can only facilitate a soft level of data integration that will not meet the standards of accuracy and completeness required by most taxonomists. Other forms of identifiers such as specimen codes and GenBank accession numbers can be used to successfully link otherwise disconnected facts in different databases, and increasingly controlled vocabularies are being constructed (e.g., GBIF-Vocabularies, 2010) to facilitate the shared use of homologous terms for this purpose. Shared identifiers also have the advantage that the structure of these links can also be exploited. For example, the PageRank algorithm (Page et al., 1999) as outlined by Page (2008) provides a means to rank search results. This addresses the problem of how to weight the many thousand of otherwise identical entries in taxonomic databases. For instance, more important specimen records (e.g. holotypes) are likely to receive more citations (links) and thus be ranked of greater importance than those with less citations.

In recognition of the importance of shared identifiers to consistently identify the same object, the biodiversity informatics community has recently invested significant effort into developing a scheme of globally unique identifiers (GUIDs) (Clark et al., 2004). Numerous methods for generating such identifiers are available with discussion primarily focussed on three alternatives (HTTP URIs, DOIs, and LSIDs); however, a clear consensus on the most appropriate for biodiversity data has yet to be reached.

Digitisation

Specimens and literature are the raw materials that feed into the scientific process of taxonomy. Enhancing access to these resources is fundamental to improving the efficiency of descriptive taxonomy, and digitisation provides a universal means of achieving this through a one-off investment. The scale of the required investment makes complete digitisation a seemingly insurmountable task (IWGSC, 2008). New technical approaches need to be considered to make this task less formidable.

It is estimated that there are more than 2.5 billion specimens in Natural History collections worldwide (Duckworth et al., 1993) and approximately 320 million pages of descriptive taxonomic literature (Hanken, 2010). For example, the Natural History Museum London alone has an estimated 70 million specimens, more than one million books, 25000 periodical titles, half a million natural history artworks with extensive map, manuscript and photographic collections and archives comprising over one million further items. Other natural history institutions have collections of a comparable size. Given the scale of these collections complete digitisation at the unit (e.g., specimen) level is almost impossible to fund. However, compared with the time, effort and money

associated with gathering information required by modern taxonomic monographs, item level digitisation may be cost effective if the processes used can ensure adequate reuse and accessibility of the digitised data. At the very least, not digitising the output from new taxonomic efforts might be seen as a missed opportunity (Krishtalka and Humphrey, 2000).

Taxonomy is the driver for most natural history digitisation efforts, and typically natural history digitisation is approached in the same rigorous, comprehensive and consequently slow manner that has become a trademark of the profession. Unfortunately the scale of the task and complex nature of many natural history items (e.g. card mounted insects held on pins) means that these approaches are manual and usually very slow. Such methods cannot even keep pace with new material entering collections, let alone a legacy of 250 years of taxonomy research. Increased mechanisation, coupled with small and often temporary compromises in the initial output from digitisation projects, can however, address this problem of scale. Combined with social incentives that better acknowledge the value of digitally published data, the task of item level natural history digitisation is possible on a manageable timeframe.

A combination of increased and more affordable computing capacity, high-quality digital cameras, and extended-focus software has made imagery of even small specimens very fast and relatively inexpensive. Crucially, when these are combined with workflows adapted to exploit the standard properties of many natural history collections, the process of digitisation can be made much faster. For example, new high-resolution photography makes it possible to image whole collection draws containing hundreds of specimens in under a minute. These images are devoid of the parallax and edge effect distortion, previously associated with composite specimen images. Because collection draws are usually of a standard size and considerably fewer than the specimens they hold (perhaps hundreds of thousands, instead of tens of millions), the complete digitisation of entire collections is possible on a reasonable timeframe. Comparable approaches can be used for items not stored in draws. For example, million of specimens are mounted on regular sized glass microscope slides. Using technology developed for imaging histological samples, slide mounted specimens can be preloaded into racks feeding scanners that can autofocus to produce high resolution images. Similarly taxonomic literature and be imaged through high throughput scanners or with imaging technology of the kind used for the BHL project. Even then, these approaches are not appropriate for all natural history collections. Some are mounted or stored in such a way that their essential characteristics cannot be imaged without item level handling (e.g. the undersides of pinned butterflies and moths, and most specimens stored in spirit). Likewise conventional images of some specimens have little or no practical value (e.g. selected palaeontological material or mineral samples where subsurface features or chemical properties are the important discriminating properties of collections). Nevertheless, in many cases some level of digitisation is both possible and of value to a major portion of the world's natural history collections. This value is significantly enhanced if images of metadata about the objects being digitised (e.g. specimen labels) can be simultaneous captured during imaging process. Even in cases where this is not possible, a unique identifier can be

assigned to each physical object and its digital image so that metadata subsequently extracted from the object can be jointly associated with both the specimen and its image.

Social issues

Sustainability

Long-term support of users, software, and the underlying hardware is arguably the greatest social challenge to new methods of publishing descriptive taxonomy. Publication of taxonomic research on the Web risks being perceived as more ephemeral than that published through traditional publication practices. Such concerns are well founded. Web links break, software and browser technologies change, users are fickle and research projects with their software developers come and go. A taxonomist considering whether to engage with new publication practices must balance such risks and the natural inertia of working with familiar tools against potential gains in efficiency, impact and personal profile derived from using novel publication approaches. At present this risk is too great for many scholars. A recent UK study of Web publishing practices across all academic sectors showed participation greatest amongst older more established scholars and younger graduate and postgraduate students, but a significant decline in participation amongst post-doctoral researchers and junior lecturers (Procter et al., 2010). These junior researchers and lecturers are highly dependent upon a cycle of traditional publications in order to maintain research grant income and enhance their academic credibility. Without engaging this large and risk adverse group of researchers, efforts to find more efficient ways of publishing are, at least in taxonomy, likely to fail.

As more taxonomists publish their data in an integrated way, these systems become increasingly valuable to other users through a growing pool of available data. These so-called “network effects” only become visible when a critical mass of users adopt the system (Benkler, 2006). Given the relatively small pool of potential taxonomists compared to other academic sectors, it is especially critical to engage and sustain mainstream taxonomic researchers in new publication techniques. Users of these systems need confidence that their contributions will be maintained and are available beyond the short lifecycle of a typical research grants. Achieving this kind of sustainability for novel forms of data publishing would traditionally have required substantial and ongoing investment in hardware, software and human capacity. However, technical advances over the past 10 years mean these costs are now substantially reduced. Hardware (the traditional capital investment) can be outsourced at relatively low cost with service level agreements that guarantee levels of access, backup and archival. In contrast software development is becoming the primary capital investment (Atkins et al., 2010). The standard scientific practice of constantly reinventing and rebuilding existing tools is increasingly untenable, leading to the potential for development of shared infrastructures that are of high enough quality to be used across multiple disciplines. These need to be designed with modern software methodologies so that they are technically sustainable. Even then, maintained support for successful projects will need sustaining beyond the lifecycle of standard research grants. Institutions that underpin taxonomic research like our major natural history national museums, might eventually consider taking on these costs for the taxonomic sector. This will only be likely if the cost of these new

publication practices is less than those of author fees and access charges to traditional publications. To make this transition, needs driven and user-led experimental projects (so called bottom-up initiatives) will need to be coupled with management (top-down) incentives to encourage use of these systems. Natural selection processes will ensure that only the fittest survive, helping to identify those projects that require sustained support outside the lifecycle of typical research grants.

Quality Control and Peer Review

Perceptions about the scholarly merit of published resources supporting taxonomic data are a key factor in the assessment and use of taxonomic research. Evidence suggests that scholars distrust novel publication processes and tools such as online databases, weblogs (blogs) and online encyclopaedia [e.g. the Encyclopedia of Life (EOL, 2010) or Wikipedia] because they fall outside the norms associated with traditional publication practices (Procter et al., 2010). In particular, ambiguity over whether a resource has been peer reviewed are central to concerns over its scholarly credibility (Nentwich, 2006). Contributors are reluctant to engage with novel publication approaches over fears that their work will be perceived as less credible and because these outputs do not (generally) count in any form of research assessment. Similarly, those using these digital resources do not have the normal cues of quality assurance (e.g. journal impact factors or peer review guarantees) to help assess scholarly value. For novel forms of publication to become mainstream, either the traditional mechanism of peer review needs to be replicated in a digital environment or other measures need to be developed to maintain confidence in taxonomic research.

Despite the dogma that peer review is the scientific community's most objective means of identifying "truth", there is surprisingly little empirical evidence for this, or even on peer reviews effectiveness in raising standards (Grayson, 2002). Recent high profile cases have highlighted flaws in the peer review process relating to instances of fraud due to fabrication, falsification, or other forms of scientific misconduct (Anonymous, 2006). These issues are particularly important to taxonomic research because descriptive taxonomy involves reporting facts that are largely impossible for a reviewer to substantiate without reference to the source taxonomic material, and because the majority of peer review is still conducted anonymously. Similarly, long running concerns over the possibility of bias or suppression by editors and reviewers have also been raised. This has even been the subject of specific study within the systematics research community (Hull, 1991). Collectively this evidence suggests that peer review is perhaps better at identifying the *acceptable truth* amongst scientists, rather than *objective truth* in science.

In practice peer review of taxonomy is mostly an effort to ensure normal taxonomic practices and standards have been observed, coupled with an assessment of the works importance relative to the perceived audience of the publication. Since the majority of descriptive taxonomic work has minimal short-term impact outside narrow specialist audiences, a subjective assessment of importance is largely irrelevant when reviewing most taxonomy. Increasingly it is common for taxonomic work to be published in taxonomic mega-journals that cross-cut many taxonomic groups (e.g. *Zootaxa* for animals, *Mycotaxon* for fungi and *IJSEM* for microbiology), leaving reviewers with the

more critical task of ensuring that authors have complied with taxonomic and editorial norms. In the light of new forms of digitised publication, it is reasonable to ask whether even these processes could be mechanised to some degree.

Automated checks of descriptive taxonomic research are conceivable because the components of published taxonomy are representations of reported facts that can be readily atomised into their component parts (Penev et al., 2009). Subjecting these components to algorithmic validation has a number of possibilities. Examples include simple checks to confirm the presence of required text (e.g. material examined) and component data (e.g. bibliographic citations, figure and table numbers and legends); validation of common data standards (e.g. geolocative data for specimens); tests to ensure consistent and positive terminology is used in diagnoses and taxonomic keys; and even tests to detect misconduct such as those used to identify plagiarism.

Arguably some of these tests already exceed the likely capabilities of human reviewers and would dictate a consistent level of quality that could be defined by the taxonomic community in collaboration with those maintaining the software infrastructure. The initial impact would be to reduce the burden on reviewers, minimise the transaction costs associated with publication and speed up the dissemination of taxonomic work. There would also be a number of benefits for the authors. Parts of the publication could be assembled automatically (e.g. maps of specimens localities from the materials examined section) and removal of human peer review would allow real time instant publication of descriptive research. The long-term effects could be even more profound. Facilitating the publication of small incremental advances in taxonomy (e.g. single species redescriptions and revisions) would facilitate more prioritised taxonomic research driven by opportunity and immediate user need. This approach would in some cases remove the necessity to produce lengthy monographs and revisions that risk never being completed / published due to changing author and publisher priorities. More importantly, the greater use of automated peer review would enable manuscript corrections to be made in versioned documents without the lengthy processes and formal procedures currently needed to correct published errors. Such errors are commonplace, especially as many traditional taxonomic journals no longer use copyeditors. Making the process of correcting errors as easy and accountable as editing a Wikipedia article would have a profound impact on the long-term quality of published taxonomic works.

Online experiments with new forms of peer review are not without precedent. Indeed, they are not even particularly new. The arXiv server (arXiv.org, 2010), founded in 1991 by physicist Paul Ginsparg is an electronic archive and distribution server primarily for physics and mathematics articles. Highly regarded within the scientific community, arXiv receives roughly 5,000 submissions each month and in 2008 passed the half million-article milestone (Ginsparg, 2008). Each article (called an e-print) is not peer reviewed but may be re-categorized by moderators. More recent attempts to address the problem of quality control in science include PLoS-One. This electronic journal covers primary research from any science and medicine discipline and employs a hybrid system in which submissions receive editorial peer review to ensure the work is rigorous but leave the wider scientific community to ascertain significance, post publication. This is achieved

through user discussion and rating features that facilitate debate and comment. More explicit two-stage reviewing is employed by the editors of *Atmospheric Chemistry and Physics* and a growing number of sister journals published by the European Geosciences Union. Editorial peer review is accompanied by a fixed term community review period during which interested parties are invited to comment. A final decision is taken at the end of this fixed term, resulting in rejection, revision or publication.

With the possible exception of arXiv, current attempts to revise scholarly peer review are evolutionary rather than revolutionary (Jefferson, 2006). So far as I am aware, nothing currently exists in scholarly circles along the lines of automated review that I have suggested for taxonomic descriptions. This is partly because the taxonomic data services needed to facilitate this process are insufficiently developed by the taxonomic community. In addition, most publishers lack the domain knowledge required to build such a specialised system. Even if they did, it is unlikely they would have sufficient commercial incentive since these systems would not simply map to other larger and more lucrative descriptive sciences (e.g. chemistry and astronomy).

A bigger challenge to addressing the failings of peer review comes from perceptions within the broader scientific community. Peer review is widely portrayed as a quasi-sacred process that makes science our most objective truth teller. Any science that rejects peer review endangers their very classification as a science. For example, peer review is a pre-condition for indexing in biomedical databases such as PubMed. For taxonomy to reject traditional peer review, it will need an exceptionally convincing case that the alternative will substantially improve quality and drive up standards.

The best case for reforms comes from the sheer volume and pace of taxonomic information enabled by the Internet and publishing tools such as blogs (Akerman, 2006). A shortage of professional (paid) taxonomists to facilitate peer review coupled with rising amateur interest and new outlets for publishing information on the Web, has the potential to create a perfect storm unless novel approaches to review can be found. On the Web, review is not so much a method of quality control, but rather a filtering and distribution system. Almost anything published (e.g. blog posts, images, videos and short articles) is discoverable thanks to various search and discovery tools that are freely available. The challenge is in filtering and aggregating this content according to its relevance and merit for a particular audience. Online, plenty of websites let their readers decide what makes front-page news. For example, Slashdot and Digg are vibrant and useful news services where users submit stories and others vote them up or down. Readers can personalise “their” front page according to their own interests, and comment on stories, providing a real-time counterpoint that is often as illuminating as the original article. These systems have a similar effect to peer review, but mostly act on content that has already been through some form of selection. Perhaps combining this form of human selection (crowdsourcing), coupled with algorithmic checks and validation as part of an integrated publishing system, offer an opportunity to genuinely challenge the primacy of traditional peer review, at least for the descriptive sciences like taxonomy.

The codes of nomenclature

The formal process of recording animals, plants, cultivated plants, prokaryotes, and viruses is governed by separate codes of nomenclature that set out rules and recommendations on how representatives of each taxonomic group should be named. The history and implementation of each code varies significantly, but their common goal is to help stabilise the naming of taxa, assisting in providing each with a unique name that is accepted worldwide. Unfortunately the absence of a unifying nomenclatural rules means that the codes governing nomenclature are implemented and interpreted differently by five separate committees. More importantly, there is no complete catalogue of all scientific names. This makes it hard to establish a name's nomenclatural validity, propagates the publication of inter-code homonyms (the same name validly published under different codes), and makes it an ordeal to establish the correct usage (concept) of a name in relation to its associated taxa (Hawksworth, 1995). These problems take on a special significance in informatics frameworks because taxonomic names are a primary means with which diverse forms of biodiversity information can be linked (see the section on *Data integration*). However, they also act to create a major social barrier in the light of changing publishing practices.

The codes of nomenclature are not legally enforceable and rely on mutual agreement and awareness for them to be implemented. Central to each is the need to define what constitutes a valid publication in order to establish priority of a taxonomic name. With the exception of viral taxa, all published works must be printed in hard copy to meet the demands of the respective codes. This excludes a growing number of journals that are only available electronically, and a potential generation of data driven publications where the constituent information is constructed from a database. Provisions controlling the acceptance of taxon names include depositing printed hardcopies of published works in a minimum number of libraries (five libraries for animals, ten for plants); reporting in a specific journal (the *International Journal of Systematic and Evolutionary Microbiology* for Prokaryotes) or registration and hardcopy printing via a committee (the *International Cultivation Registration Authority* for plant cultivars). Viral taxa are handled in a similar way to cultivars, in that proposals must be approved by a committee before formal recognition, and this requires sufficient characterisation in the published literature. There is, however, no requirement that these published works be printed in hardcopy. Electronic-only publishers of nomenclatural acts relating to animals and non-cultivated plants can remain compliant with the respective codes by print and mailing hard copies of articles to the minimum number of libraries. But the high number of nomenclatural acts (circa 24-30 thousand each year for Zoology alone, Chapman, 2009) and the growing number of publishers moving to electronic only publication (EPS-Ltd., 2006), means that this is not a scalable solution. To date this approach has been largely restricted to high profiles case (e.g. *Darwinius masillae*, a primate fossil published in *PLoS-One*), where electronic publication initially fell foul of the nomenclatural codes. The demise of print publishing means events like these are likely to become more common, and because the nomenclatural codes are not mandatory there is an increasing risk that provisions of the code will be ignored.

In the short term there is an urgent need to reform the respective nomenclatural codes to accept electronic-only publications while taking reasonable steps to ensure the permanency of the original article. This is particularly important for nomenclatural acts relating to animal and plant names, since prokaryotes, viruses, and plant cultivars already have the equivalent of a central nomenclatural register and account for tiny fraction of all nomenclatural acts. Such an amendment is currently under consideration by the Zoological Commission regulating animal names (ICZN, 2008), and is supported by the ongoing development of ZooBank, an official online registry of zoological nomenclature (Pyle and Michel, 2008). ZooBank has the potential to ensure that electronic-only publications remain accessible and unchanged, even if the original article cannot be accessed. However, whether ZooBank takes on this role is a matter of ongoing discussion (Polaszek et al., 2008). For plant names, the likelihood of an equivalent amendment to the botanical code looks bleak. At the XVII International Botanical Congress held in Vienna in 2005, the Special Committees on Electronic Publication had both its (alternative) proposals to facilitate the acceptance of electronic-only published acts rejected. These amendments received more than 75% "No" votes in a mail ballot (McNeill, 2006). Arguably the Botanical commission had not done enough to ensure the permanency of electronically published nomenclatural acts. This problem needed to be convincingly addressed before attempts to facilitate electronic-only publishing are put before the zoological community, or resubmitted to the botanical community.

In the longer term more radical reform is needed to facilitate a common nomenclatural code for all forms of life. Based on a survey of UK taxonomists Hawksworth (1992) estimated that approximately 20% of a taxonomists time is spent on nomenclatural work, and suggested that this cost the equivalent of 25-50 taxonomists posts in the UK alone. These taxonomists might otherwise be working on the biology of the organisms, if they were not working on this system of labelling. Our increasing reliance on electronic information retrieval systems further intensifies the case for an unambiguous means to label and identify all forms of life. Without unique identifiers, much of the information accumulated by past and present biodiversity research is effectively irretrievable.

A recent attempt to unify the five nomenclatural codes culminated in the draft BioCode (Hawksworth et al., 1996). This showed that a single, simpler code of nomenclature for all forms of life is feasible. The BioCode was modelled on the approach taken by the Bacteriological Code of stabilising names by approved lists, coupled with a new simple code for the nomenclature of the future that synthesized elements of the existing codes. However, the BioCode draft has received little attention since 1997; its originally planned implementation date of January 1, 2000, has passed unnoticed. Another unified code in development is the PhyloCode, which aims to regulate a phylogenetic nomenclature of all life (Queiroz and Cantino, 2001). Implementation is tentatively scheduled for sometime before 2010, but arguably the PhyloCode even more doomed than the BioCode due to the instability and incompleteness of phylogenetic hypotheses. Both the BioCode and the PhyloCode are naive to the sociological challenges of overturning systems of nomenclature whose foundations were laid more than two hundred and fifty years ago.

Biological nomenclature is not an end in itself. Arguably it is not even a scientific endeavour. There is a profound need for a more automated assignment of unique identifiers for all forms of life. Such a scheme should not discriminate in the assignment of different types of identifiers to biological objects and can sit alongside traditional nomenclatural schemes until it has proven its worth. An increasing numbers of biological samples are already circumventing traditional taxonomic nomenclature due to the lack of resources, time and expertise on the part of the taxonomic community. This is a particular challenge for the DNA barcoding community. A growing number of taxonomic groups are entering this post-taxonomic era when traditional taxonomy is not tenable due to the volume of data being produced and a lack of expertise. It would be good if we entered this period with a unified system of labelling that did not have the problems inherent to traditional codes of nomenclature and facilitated the open sharing and preservation of these identifiers to address the technical challenges of integrating diverse forms of biodiversity information.

Copyright and intellectual property

Technology had radically altered the way taxonomists and the wider scientific community generate, access and publish information. In consequence scientists cannot help but come into contact with copyright law and intellectual property issues through everyday activities like browsing the Web, or publishing some text. Technology, heedless of copyright law, has developed modes that insert multiple acts of reproduction and transmission that are actionable events under copyright statutes (Boyle, 2009). Unchallenged, current copyright legislation is counterproductive for science, even in the parochial world of descriptive taxonomy.

Copyright law was developed to regulate the creation of copies. In an analogue world of books and other printed resources this meant that many uses of analogue works are unregulated and free. For example, the act of reading, lending or selling a book is exempt from copyright law because these ordinary acts do not create a copy. In contrast in the digital world almost every single use of a digital object creates copies that potentially triggers the reaction of copyright. This has occurred, not because legislators or politicians have changes the law. It has occurred because the platform through which people gain access and create these works has changed. In short, copyright law was developed for a radically different objective and is being unthinkingly applied across a range of contexts that were never originally intended (Lessig, 2008).

The paradigm case in copyright law concerns certain professionals who depend upon an exclusive right to control copies of their work as part of their business model. Copyright is a means by which these professionals (e.g. musicians) exercise their claim to this exclusivity in order to secure profit. Copyright was not developed to address the needs of other communities that rely on different business models to sustain their activities. For example, most of the scholarly community secure funds as a condition of employment though publicly funded academic institutions, and not (usually) through controlling access to the products of work. On the contrary, scholars usually seek the widest possible dissemination of their work, incentivised by the peer recognition they receive and (other things being equal) commensurate financial remuneration through their employer. This

was recognised by the famous sociologist of science Robert Merton who wrote about the common ownership of scientific discoveries, through which scientists give up their intellectual property rights (including copyright) in exchange for peer recognition and esteem. Merton (1942) said that “property rights in science are whittled down to a bare minimum by the rational of the scientific ethic.” A similar argument can be made for amateur communities, e.g. amateur taxonomists, who incidentally significantly outnumber professional taxonomists, that by definition rely on a different business model to sustain their amateur interests.

Copyright, as the exclusive right to make copies, is harmful to the scientific ethic but is used successfully by publishers to incentivise their publishing efforts. This is achieved through the profit gained by publishers controlling access to their printed works. The tension created by this between publishers and scientists has led to the development of new business models for science publishing in which the author, and not the consumer, pays the publisher for disseminating works (Velterop, 2004). This model, termed Open Access, has been independently accompanied by the development of licences (notably the Creative Commons licences, 2010). The combination of Open Access publishing and Creative Commons licences enables users to gain free access and reuse of content under the specific terms of the licence without having to refer back to the rights holder. To date approximately 1,000 scholarly journals have adopted Open Access (EPS, 2006), although this is often withheld for a period so that recent journal content can be sold in the traditional way. In taxonomy just one publisher (Pensoft) has adopted wholesale Open Access by charging a modest fees. Arguably this is because the economics of taxonomic research mean that most taxonomists (especially amateur taxonomists) do not have access to the funds required to pay the substantial fees of mainstream Open Access journals.

The problem of copyright is particularly profound for taxonomy, because taxonomic works are often multi-authored compilations that contain substantial elements drawn from other published work. For example, a taxonomic monograph may contain an identification key with illustrations sourced from many hundreds of different publications. Under current copyright law reproduction of these illustrations requires tracing the rights and securing permissions from hundreds (possibly thousands) of publishers who retain the image rights. This expensive and time-consuming process is often impossible because rights holders cannot be traced. In these cases the scholar must prove they have performed due diligence when attempting to tracing the rights to these so-called Orphaned works, before they can be used (DCMS-&-BIS, 2009). The situation is further compounded by the fact that property rights can extend beyond traditionally published works to facts held in databases. In the USA facts known in the public domain but retrieved from a database can retain copyright protection if the arrangement of the data is deemed to be “sufficiently original”. In the EU copyright protection only extends to databases if their contents are considered to be "original literary works" (e.g. original work which is written, spoken or sung), but additional protection is automatically conferred to databases if "there has been a substantial investment in obtaining, verifying or presenting the contents of the database" (further details in the UK *Copyright and Rights in Databases Regulations Act*, 1997). As such, databases of taxonomic facts such as the *Species 2000* checklist of taxonomic names (Species-2000, 2010), and the

Thompson Reuters *Index to Organism Names* (ION, 2010), assert these rights of protection (e.g. Harling and Bisby, 2004), hindering reuse of their factual contents and setting a precedent that encourages other taxonomic database creators to use the same restrictive practices. An unfortunate side effect of the rights problem in taxonomy is that some rights holders seek to aggressively brand derivative works containing elements of their products in an attempt to maintain the independent profile of their projects. For example, aggregation efforts like the *Encyclopedia of Life* are littered with logos of contributing projects, diminishing the incentive for fresh contributions.

The original copyright legislation was not conceived for future technical opportunities afforded by new technologies such as the Internet and databases in scholarly research. Consequently our copyright legislation has grown into a patchwork of technical, inconsistent and complex rules that touch almost every act of the scholarly process (Lessig, 2008), even in parochial disciplines like taxonomy. This regime is blocking our legal access to mix and reuse scholarly research like taxonomic data. For taxonomy our primary defence is that no one sufficiently cares to take action against our community. This is (usually) because there is little or no financial loss to the affected parties. Nevertheless, it would be dangerous to build a new paradigm for database driven taxonomy on a foundation that is illegal. Therefore we need to find ways to avoid these legal barriers for taxonomy to progress.

Wider use of Creative Commons licences are central to this process. These enable authors of content to legally define the freedoms they give in ways that can be read by ordinary people, lawyers (so they are legally enforceable), and computers through expressions written in computer code (specifically Resource Description Framework). This has been extended for science through the Science Commons deeds (Science-Commons, 2010) that provide a legal framework for sharing science data, including physical objects through standardised material transfer agreements. As with the Creative Commons licences these deeds have three layers of access providing versions that can be read and understood by scientists, lawyers and computers. The goals of this effort are to facilitate the scientific ethic of sharing through an efficient intellectual property framework. This addresses many of the unintended consequences of our outdated copyright legislation associated with changing technologies for the production and dissemination of scientific information.

Early solutions to e-publishing descriptive taxonomy

New technologies have created an opportunity to revolutionise how taxonomy is performed and shared. These lend themselves to new forms of publishing that transcend the traditional linear flow of information from author to reader. Taxonomists are beginning to experiment with Virtual Research Environments (VRE's) that enable real-time collaboration and dissemination of taxonomic information while simultaneously affording new opportunities for data stewardship, curation, and data mining. These environments, while highly experimental, offer an alternative vision for the practice of taxonomic research that challenges the purpose and definition of traditional taxonomic publishing.

The following review of VRE's is not intended to be comprehensive. Rather it is intended to be illustrative of the diverse approaches taken to tackling these challenges. Most of these projects are experimental and suffer from problems of stability and sustainability inherent to new software development. Notably each has been developed within the taxonomic community, rather than via independent commercial publishers. This reflects the specialised needs and purposes of these environments as seen by the taxonomic community, rather than a from a traditional publishers perspective.

Scratchpads

Scratchpads (Smith et al., 2009, Scratchpads, 2010) were developed through EU funded EDIT project (EDIT, 2010) as a data-publishing framework for groups of people to create their own virtual research communities supporting natural history science. These cater to the particular needs of individual research communities through a common database and system architecture. This is based on the Drupal content management system (Drupal, 2010) and provides a scalable, flexible resource that facilitates the collaboration of distributed communities of taxonomists. Sites are hosted, archived and backed up by the Natural History Museum London and are offered free to any scientist that completes an online registration form. Registrants assume responsibility for the contents of each site, which (on approval) are instantiated at Web domains of their choice. The default Scratchpad template has workflows to support the upload and publication of bibliographies, image galleries, specimen records, character data sets, documents, maps and custom data defined by the contributor. These can be uploaded to the site *en masse* through spreadsheets or created *de novo* through editing interfaces that support entry for single and multiple data records. Data are classified and aggregated around a taxonomy supplied by the user or imported automatically from a service provided by the Encyclopedia of Life project. Authenticated users can optionally supplement these data with information drawn from high quality Web accessible databases (e.g. GBIF and the Biodiversity Heritage Library). This facilitates the rapid construction, curation and publication of content rich Web pages about any taxon. Users can withhold public access to unregistered users or create private groups. However, all public content is published by default under a Creative Commons (by-nc-sa) license, which is a condition of use for all site contributions except multimedia. In addition, selected data types (taxonomy, specimens and literature) can be accessed through Web services using documented standards and protocols. High-level administrative functions including the control of permissions, user roles, and the development of new functionality, are centrally managed by the Scratchpad development team. However site contents and access rights are owned and controlled by registered users with sufficient privileges as dictated by the site maintainers.

The Scratchpad framework currently serves more than 1,800 registered users across more than 160 sites, spanning scientific, amateur and citizen science audiences. Sites range in function from supporting the work of societies and conservation efforts to the production and dissemination of taxonomic checklists, peer reviewed journal articles and electronic books. As a derivative of Drupal, the source code is covered by version 3 of the GNU General Public License and is available from a subversion repository (Scratchpad-SVN, 2010).

MX (MatriX)

MX is a Web-based content management system that is conceptually analogous to the Scratchpad project. MX enables taxonomists working at multiple institutions to access the same shared information and data to build and publish taxonomic and phylogenetic research. The system can be used for the collaborative coding, storage and manipulation of multiple data types (morphological, molecular, specimen, image, descriptive, bibliographic, label, collection, and associated information) through a grid matrix style editor. A series of tabs provide a simple workflow that allows the user to navigate these data types. These are connected to engines that dynamically publish content in various forms. These include multiple-entry and dichotomous on-line keys with taxonomic catalogues (e.g., MX-Keys, 2010) associated host data (Diapriidae, 2010); linked images from the MorphBank image database (Morphbank, 2010) and supports various data export formats. MX is also integrated with a simple ontology builder (Hymenoptera-Glossary, 2010), such that controlled terms in blocks of descriptive text (e.g. anatomical characters) can automatically be linked to referenced and illustrated term definitions.

MX was developed by Matt Yoder and Krishna Dole as part of NSF funded projects at Texas A&M University. It is an open source (MySQL, Apache, Ruby on Rails) application that is principally used by researchers working on US funded Hymenoptera systematics projects.

CATE

CATE (Creating A Taxonomic E-science) (CATE, 2010, Godfray et al., 2007) is a Web-based system developed to facilitate the construction of consensus or consolidated taxonomic hypotheses. These are presented and navigated through a taxonomic hierarchy containing data common to modern taxonomic monographs (e.g. nomenclatural information, descriptive data, specimen records, observations, images, molecular data). The taxonomic hierarchy and associated content is added through an online workflow that supports peer review, enabling authorities with editorial privileges to comment and amend contributions as required. CATE facilitates an incremental cycle of revision and publication, with the current consensus classification and alternative taxonomic hypotheses being presented to end-users, but with earlier versions and withdrawn hypotheses preserved and archived. Contributions (referred to as “proposals”) are refereed and opinions sought from the taxonomic community before a committee decides whether they should be incorporated into the next edition of the consensus taxonomy. As such, CATE moves beyond the paradigm of a static publication, providing a means for taxonomists to revise and publish the latest taxonomy of a group with traditional peer review, while documenting and archiving previous incremental changes.

CATE was established for two model groups (plants belonging to the Arum family [<http://www.cate-araceae.org/>] and Sphingid hawkmoths [<http://www.cate-sphingidae.org/>]) in order to ensure that the system is compliant with the Botanical and Zoological codes of nomenclature. A significant component of the CATE project concerned the digitisation of specimens and literature relevant to the model taxa. Technically CATE is a Java Web application based upon the Spring Web Flow framework (Spring, 2010). Data are stored in a relational database (MySQL, 2010) that

was developed in conjunction with the EDIT Common Data Model (see the *EDIT Cyberplatform*). The project is currently restricted to the two model organism groups. However, the principles developed within CATE are intended to be exploited as part of a larger initiative led by Kew Gardens to develop a system supporting the construction of an online revision for all monocot plants.

EDIT Cyberplatform

The EDIT Platform for Cybertaxonomy (Cyberplatform, 2010) is a collection of tools and services that integrate to cover various aspects of the taxonomic workflow. The workflow is grouped into various editing activities (taxonomic hierarchies, collection and specimen records, descriptions, documenting fieldwork, literature management, and managing GIS data) culminating in tools to generate taxonomic manuscripts for publication. At the heart of the Cybertaxonomy platform is a database (the Common Data Model - CDM), which acts as a repository for data produced by individual or groups of taxonomists in the course of their work. The CDM also acts as a technical back-end for data services that can be accessed by developers through a software (Java) library. The primary Cybertaxonomy platform components consist of a desktop taxonomic editor, software to build a data portal, the CDM Java Library, a specimen search portal plus various tools supporting GIS related functions. This software is packaged in three forms to enable individuals, local research groups and distributed research groups to install respectively on a desktop computer, a local intranet or on the Internet. At present the Cyberplatform is primarily by used by three exemplar groups that are explicitly funded through the European Distributed Institute of Taxonomy to use this system, although selected components of the Cyberplatform are used more widely. Notably the CDM is used by several taxonomic communities including the European Register of Marine Species (ERMS, 2010) and the Euro+Med Plantbase project (Euro+Med, 2010).

Solanaceae Source

Solanaceae Source (Solanaceae-Source, 2010) is taxonomic database for members of the genus *Solanum* that has been published on the Web. The treatment is the result of the PBI Solanum project, a worldwide study funded by the National Science Foundation under the Planetary Biodiversity Inventory program. Nomenclatural data (e.g. type records, authors and publications) coupled with additional descriptive information, illustrations and keys have been sourced from existing modern monographs. These have been coupled with new work coordinated on the remaining groups to build a definitive Web-based guide to Solanaceae. The IT component of this project differs from others listed here in that the development of a generic IT infrastructure capable of supporting other taxonomic projects was not the primary goal of Solanaceae Source. Instead the project makes use of two botanical databases (BRAHMS at the Natural History Museum London and KeEmu at New York Botanical Gardens) to import and manage the information. These data are aggregated into a customised relational model database server (MSSQL) and presented on the Web in the form of taxonomic treatments. The goal is to use these tools to produce printed manuscripts of new taxonomic research in addition to a comprehensive Web resource on Solanum.

Species File

Species File (SpeciesFile, 2010) is a collection of programs that provides access to, and manipulation of, taxonomic information stored in multiple databases. Each database provides detailed information about taxa contained within the scope of the "apex taxon" and are coupled to a website that provides the means to interact with and modify information held within the database. Central to Species File functions is the means to store nomenclatural information in a way that is fully compliant with the codes of Zoological nomenclature. The Species File development team has produced a template that contains the basic database table structure as well as the stored procedures, user-defined functions and views used by independent Species Files. The template is used as the starting point for new species file databases and is capable of a limited number of customisations.

Species File has been developed by David Eades and colleagues in conjunction with the Illinois Natural History Survey. It was conceived and developed to manage information on the insect order Orthoptera (Orthoptera-SpeciesFile, 2010), but in recent years has been adopted by ten other insect research communities. At the time of writing Species File databases contain taxonomic information on almost 52,000 valid species (circa 85,000 names) of which roughly half come from Orthoptera Species File.

The Species File Group (SFG) has developed a template that contains the basic database table structure as well as the stored procedures, user-defined functions and views used by Species File's. The template is used as the starting point for new Species File databases (SpeciesFile-Databases, 2010). Species File is a Visual Studio application. Programming for Species File is done with Visual Studio.NET. The Active Server Pages (ASP) use Visual Basic Script and client side programming is done using JavaScript.

The future of descriptive taxonomic publishing

"We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices"

E. O. Wilson | Harvard University

What is startling about many ongoing efforts within the practice of taxonomy is that they are not more effectively integrated. This generates redundancy as stakeholders both inside and outside the taxonomic community invest significant time discovering, re-organising, integrating and analysing these resources, usually at great expense. The examples of Virtual Research Environments highlighted here suggest a way forward that addresses many of these challenges. These offer a vision for not only automating aspects of publishing taxonomy, but also to apply new methods that have the potential to revolutionise how taxonomic research is practised. Such systems, with sufficient users, could enable the tackling of grand challenge problems that are untenable by other means, because of the opportunities they create for large scale stewardship, curation, and mining of enormous collections of heterogeneous taxonomic data. Online observatories that engage new users for taxonomy are also conceivable, thanks to the ease with which

taxonomic data can be repurposed and represented for different audiences from a database. Likewise, more rational development and use of research instrumentation could be planned, since this can be plugged into VRE's to dramatically reduce barriers of time and distance (distance in the geographic, disciplinary, and organisational sense) that would otherwise interfere with the construction and operation of these systems. In the near term VRE's for taxonomy might integrate with analytical tools that require enormous storage and computing capacity. For example, in the morphometric analysis of specimen images for automated identification, or the analysis of very large phylogenetic trees. In the longer term, VRE's may be connected directly to research collections in major museums and herbaria, providing remote to specimens for conducting taxonomic research.

Central to the transformation of descriptive taxonomy is the need to change our concept of publication. In a database the power of taxonomy is amplified by ingenuity through applications and uses unimagined by the original authors and distant from the original field. In a paper, taxonomy languishes in obscurity, inaccessible and unused by all but the most determined. At present publicly funded research requires "classical" publications. These attract peer recognition that influences the authors' reputation, employment and research opportunities. Without expanding the concept and recognition of publications that embrace alternative forms of dissemination, the chances of transforming taxonomy are significantly diminished. Through tools like VRE's the Web can be used as an instrument of scientific research, blending social and technical solutions to address challenges that are otherwise insurmountable. As these tools mature taxonomists, for their discipline to survive, must embrace them.

Acknowledgements

Thanks to Chris Lyal for inviting me to write this chapter and Dave Roberts for providing valuable comments and corrections.

References

- ABCD (2010) *Access to Biological Collections Data (ABCD)*. Accessed 1 March, 2010 from <http://www.tdwg.org/activities/abcd/>.
- Akerman, R. (2006) Evolving peer review for the Internet. *Nature (online edition)*, doi:10.1038/nature04997.
- Anonymous (2006) Peer review and fraud. *Nature*, 444, 971-972.
- arXiv.org (2010) *arXiv.org e-print service*. Accessed 1 March 2010 from <http://arxiv.org/>.
- Atkins, D., Borgman, C., Bindoff, N., *et al.* (2010) Building a UK Foundation for the Transformative Enhancement of Research and Innovation. *Report of the International Panel for the 2009 Review of the UK Research Councils e-Science Programme*. Research Councils UK.
- Benkler, Y. (2006) *The Wealth of Networks*, New Haven and London, Yale University Press.
- BHL (2010) *Biodiversity Heritage Library*. Accessed 1 March 2010 from <http://www.biodiversitylibrary.org/>.

- Bourne, P. (2005) Will a Biological Database Be Different from a Biological Journal? . *PLoS Computational Biology*, 1, e34.
- Boyle, J. (2009) *The Public Domain: Enclosing the Commons of the Mind*, Yale University Press
- Brown, D. J. (2008) *The impact of electronic publishing: the future for libraries and publishers* K G Saur Verlag.
- CATE (2010) *Creating a Taxonomic e-Science*. Accessed 1 March 2010 from <http://www.cate-project.org/>.
- Chapman, A. D. (2009) Numbers of Living Species in Australia and the World. *Report for the Australian Biological Resources Study*. Canberra, Australia.
- Clark, T., Martin, S. & Liefeld, T. (2004) Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, 5, 59-70.
- Constable, H., Guralnick, R., Wieczorek, J., Spencer, C. & Peterson, A. T. (2010) VertNet: A New Model for Biodiversity Data Sharing. *PLoS Biology*, 8, e1000309.
- Creative-Commons (2010) *Creative Commons*. Accessed 1 March 2010 from <http://creativecommons.org/>.
- Cyberplatform (2010) *EDIT Cyberplatform*. Accessed 1 March 2010 from <http://wp5.e-taxonomy.eu/>.
- Dallwitz, M. J. (1980) A general system for coding taxonomic descriptions. *Taxon*, 29, 41-46.
- Darwin-Core (2010) *Darwin Core Schemas*. Accessed 1 March 2010 from <http://wiki.tdwg.org/twiki/bin/view/DarwinCore/DarwinCoreVersions>.
- DCMS-&-BIS (2009) *Digital Britain*, London, The Stationery Office.
- Diapriidae (2010) *The Diapriidae*. Accessed 1 March 2010 from <http://www.diapriid.org/public/site/diapriid/home>.
- Drupal (2010) *Drupal content management system*. Accessed 1 March 2010 from <http://drupal.org/>.
- Duckworth, W. D., Genoways, H. H. & Rose, C. L. (1993) *Preserving Natural Science Collections: Chronicle of Our Environmental Heritage*. Washington, D.C., National Institute for the Conservation of Cultural Property.
- EDIT (2010) *European Distributed Institute of Taxonomy*. Accessed 1 March 2010 from <http://e-taxonomy.eu/>.
- EOL (2010) *Encyclopedia of Life*. Accessed 1 March 2010 from <http://www.eol.org/>.
- EPS-Ltd. (2006) UK scholarly journals 2006 baseline report: an evidence-based analysis of data concerning scholarly journal publishing. Prepared on behalf of the Research Information Network, Research Councils UK and the Department of Trade & Industry.
- ERMS (2010) *European Register of Marine Species*. Accessed 1 March 2010 from <http://www.marbef.org/data/erms.php>.
- Euro+Med (2010) *Euro+Med PlantBase*. Accessed 1 March 2010 from <http://www.emplantbase.org/>.
- GBIF (2010) *Global Biodiversity Information Facility*. Accessed 1 March 2010 from <http://www.gbif.org/>.
- GBIF-Statistics (2010) *Global Biodiversity Information Facility Data Sets*. Accessed 1 March 2010 from <http://www.gbif.org/>.

- GBIF-Vocabularies (2010) *GBIF Vocabularies*. Accessed 1 March 2010 from <http://vocabularies.gbif.org/>.
- Ginsparg, P. (2008) The global-village pioneers. *Physics World*.
- Godfray, H. C. J., Clark, B. R., Kitching, I. J., Mayo, S. J. & Scoble, M. J. (2007) The Web and the structure of taxonomy. *Systematic Biology*, 56, 943-955.
- Grayson, L. (2002) Evidence based policy and the quality of evidence: rethinking peer review. ESRC UK Centre for Evidence Based Policy and Practice.
- Hall, M. B. (2002) *Henry Oldenburg: Shaping the Royal Society*, OUP Oxford.
- Hanken, J. (2010) The Encyclopedia of life: a new digital resource for taxonomy. IN Polaszek, A. (Ed.) *Systema Naturae 250 - The Linnaean Ark*. CRC Press.
- Harling, P. & Bisby, F. (2004) Species 2000 Europa IPR Licence and Access Agreements.
- Hawksworth, D. L. (1992) The need for a more effective biological nomenclature for the 21st century. *Botanical Journal of the Linnean Society*, 109, 543-567.
- Hawksworth, D. L. (1995) Steps along the road to a harmonized bionomenclature. *Taxon*, 44, 447-456.
- Hawksworth, D. L., Greuter, W., McNeill, J., *et al.* (1996) Draft BioCode: The prospective international rules for the scientific names of organisms. *Bulletin of Zoological Nomenclature*, 53, 148-166.
- Hull, D. L. (1991) *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*, Chicago, University of Chicago Press.
- Hymenoptera-Glossary (2010) *Hymenoptera Glossary*. Accessed 1 March 2010 from <http://hymglossary.tamu.edu/>.
- ICZN (2008) Proposed amendment of the International Code of Zoological Nomenclature to expand and refine methods of publication. *Zootaxa*, 1908, 57-67.
- ION (2010) *Index to organism names*. Accessed 1 March 2010 from <http://www.organismnames.com/>.
- IWGSC (2008) Mission-Critical Infrastructure for Federal Science Agencies. A report by the Interagency Working Group on Scientific Collections. National Science and Technology Council.
- Jefferson, T. (2006) Models of quality control for scientific research. *Nature (online edition)*, doi:10.1038/nature05031.
- Johnson, N. F. (2010) Future taxonomy today: new tools applied to accelerate the taxonomic process. IN Polaszek, A. (Ed.) *Systema Naturae 250 - The Linnaean Ark*. CRC Press.
- Krishtalka, L. & Humphrey, P. S. (2000) Can Natural History Museums Capture the Future? *BioScience*, 50, 611-617.
- Lessig, L. (2008) *Remix - Making Art and Commerce Thrive in the Hybrid Economy*, Bloomsbury Academic.
- McNeill, J. (2006) XVII International Botanical Congress: summary report of the actions of the nomenclatural section of the congress - Vienna, Austria 12-16 July, 2005. *Botanical Electronic News*, 356.
- Mediawiki (2007) *MediaWiki, The Free Wiki Engine*. Accessed from <http://www.mediawiki.org/>.
- Merton, R. (1942) Science and Technology in a Democratic Order. *Journal of Legal and Political Sociology*, 1, 115-126.

- Morphbank (2010) *Morphbank*. Accessed 1 March 2010 from <http://www.morphbank.net/>.
- Mx (2010) *MX (MatriX) Software*. Accessed 1 March 2010 from <http://sourceforge.net/projects/mx-database/>.
- Mx-Keys (2010) *MX Keys*. Accessed 1 March 2010 from <http://hymenoptera.tamu.edu/keys/>.
- MySQL (2010) *MySQL*. Accessed 1 March 2010 from <http://www.mysql.com/>.
- Nentwich, M. (2006) Cyberinfrastructure for next generation scholarly publishing. IN Hine, C. M. (Ed.) *New infrastructures for knowledge production*. London, Information Science Publishing.
- Orthoptera-Speciesfile (2010) *Orthoptera Species File*. Accessed 1 March 2010 from <http://osf2.orthoptera.org/>.
- Page, L., Brin, S., Motwani, R. & Winograd, T. (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.
- Page, R. D. M. (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, 9, 345-354.
- Penev, L., Erwin, T., Miller, J., *et al.* (2009) Publication and dissemination of datasets in taxonomy: ZooKeys working example. *Zookeys*, 11, 1-8.
- Polaszek, A., Pyle, R. & Yanega, D. (2008) Animal names for all: ICZN, ZooBank, and the New Taxonomy. IN Wheeler, Q. D. (Ed.) *The New Taxonomy*. Boca Raton, CRC Press.
- Procter, R., Williams, R. & Steward, J. (2010) Use and Relevance of Web2.0 Resources for Researchers. London, Research Information Network.
- Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37, D32-6.
- Pyle, R. L. & Michel, E. (2008) ZooBank: Developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa*, 1950, 39-50.
- Queiroz, K. & Cantino, P. D. (2001) Phylogenetic nomenclature and the PhyloCode. *Bulletin of Zoological Nomenclature*, 58, 254-270.
- Rinaldo, C. (2009) The Biodiversity Heritage Library: Exposing the Taxonomic Literature. *Journal of Agricultural & Food Information*, 10, 259 - 265.
- Science-Commons (2010) *Science Commons*. Accessed 1 March 2010 from <http://sciencecommons.org/>.
- Scratchpad-SVN (2010) *Scratchpads SVN Repository*. Accessed 1 March 2010 from <http://svn.scratchpads.eu/viewvc/scratchpads/>.
- Scratchpads (2010) *Scratchpads*. Accessed 1 March 2010 from <http://scratchpads.eu/>.
- Smith, V. S., Rycroft, S. D., Harman, K. T., Scott, B. & Roberts, D. (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life *BMC Bioinformatics*, 10(Suppl 14): S6, doi:10.1186/1471-2105-10-S14-S6.
- Solanaceae-Source (2010) *Solanaceae Source*. Accessed 1 March 2010 from <http://www.nhm.ac.uk/research-curation/research/projects/solanaceaesource/>.
- Species-2000 (2010) *Species 2000*. Accessed 1 March 2010 from <http://www.sp2000.org/>.

- Speciesfile (2010) *Species File*. Accessed 1 March 2010 from <http://software.speciesfile.org/>.
- Speciesfile-Databases (2010) *Species File Databases*. Accessed 1 March 2010 from <http://software.speciesfile.org/Files/Files.aspx>.
- Spring (2010) *Spring Web flow framework*. Accessed 1 March 2010 from <http://www.springsource.org/webflow>.
- TDWG-Statistics (2010) *TDWG Biodiversity Information Projects of the World*. Accessed 1 March 2010 from <http://www.tdwg.org/about-tdwg/>.
- Velterop, J. (2004) Open Access: Science Publishing as Science Publishing Should Be. *Serials Review*, 30, 308-309.
- Walter, D. E. & Winterton, S. (2007) Keys and the Crisis in Taxonomy: Extinction or Reinvention? *Annual Review of Entomology*, 52, 193-208.