Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Taxonomy as an eScience

Benjamin R Clark, H. Charles J Godfray, Ian J Kitching, Simon J Mayo and Malcolm J Scoble

Phil. Trans. R. Soc. A 2009 **367**, 953-966
doi: 10.1098/rsta.2008.0190

References This article cites 26 articles, 3 of which can be accessed free
http://rsta.royalsocietypublishing.org/content/367/1890/953.full.html#ref-list-1

EXiS Open Choice This article is free to access

Subject collections Articles on similar topics can be found in the following collections

computational biology (59 articles)

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions

This journal is © 2009 The Royal Society

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 953–966
doi:10.1098/rsta.2008.0190
Published online 16 December 2008

Taxonomy as an eScience

By Benjamin R. Clark[1], H. Charles J. Godfray[1], Ian J. Kitching[2], Simon J. Mayo[3] and Malcolm J. Scoble[2,*]

[1]*Department of Zoology, University of Oxford, Oxford OX1 3PS, UK*
[2]*Department of Entomology, Natural History Museum, London SW7 5BD, UK*
[3]*Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK*

The Internet has the potential to provide wider access to biological taxonomy, the knowledge base of which is currently fragmented across a large number of ink-on-paper publications dating from the middle of the eighteenth century. A system (the CATE project) is proposed in which consensus or consolidated taxonomies are presented in the form of Web-based revisions. The workflow is designed to allow the community to offer, online, additions and taxonomic changes ('proposals') to the consolidated taxonomies (e.g. new species and synonymies). A means of quality control in the form of online peer review as part of the editorial process is also included in the workflow. The CATE system rests on taxonomic expertise and judgement, rather than using aggregation technology to accumulate taxonomic information from across the Web. The CATE application and its system and architecture are described in the context of the wider aims and purpose of the project.

Keywords: Web-based taxonomy; online peer review; taxonomic revisions

1. Introduction

Biological taxonomy (or systematics) is a broad field. It includes, *inter alia*, the description of species (and other taxa), the production of means of identification (such as keys), the resolution of nomenclature and phylogenetic reconstruction. Taxonomy has a long history, having becoming formalized in the middle of the eighteenth century when Carl von Linné (Carolus Linnaeus) established the binomial system of nomenclature (Linnaeus 1753, 1758), whereby a species received a double name—the species name (e.g. *catus* for the domestic cat) and the genus name (*Felis*). Combined, these names form the Linnaean binomial *Felis catus*. The protocol continues to this day. Since the time of Linnaeus, the number of species discovered and described has grown to approximately 1.8 million (Stork 1993). Just as species (e.g. *catus*) are grouped into genera (e.g. *Felis*), so genera are grouped into families (e.g. Felidae) and so on up a nested hierarchy of taxa, notably orders (e.g. Carnivora), classes (e.g. Mammalia) and phyla (e.g. Chordata). Resolving the phylogenetic relationships of species and higher taxa is a major interest of taxonomists and involves close study of anatomical, molecular and other kinds of data. Yet what has emerged

* Author for correspondence (m.scoble@nhm.ac.uk).

One contribution of 24 to a Discussion Meeting Issue 'The environmental eScience revolution'.

This journal is © 2008 The Royal Society

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

as a pressing problem in taxonomy, given the loss of biodiversity, is making available the wealth of information about species gathered through the ages (e.g. Godfray & Knapp (2004) and references therein). This contribution describes a system proposed to help solve what has become an information impediment about our basic knowledge of biodiversity.

The most inclusive compendiums of taxonomic information take the form of monographs, faunas and floras and taxonomic 'revisions'. Such works are typically comprehensive for the taxon or geographical area in question. What distinguishes such compilations is their highly contextual and synthetic content. Compared with the number of known species, there are, unfortunately, few such treatments (e.g. Wheeler *et al.* 2004). This is unsurprising as such monographs may take years to produce, requiring the gathering and careful study of specimens, which are often scattered across collections in many institutions in several countries. An appropriate digestion and synthesis of the information that exists in the published literature, which may date back to the time of Linnaeus, adds considerably to the time taken to complete these publications. Furthermore, the tendency of such works to be lengthy and heavily illustrated, ideally with many figures in colour, means that monographs typically are costly to print and distribute (Minelli 2003). While a good monograph is long lasting for the synthesis it offers, it suffers the serious limitation of conventional (ink-on-paper) publication in that its content becomes out of date as soon as it appears: new species are found; distributions change with habitat loss and the discovery of new records; and new phylogenetic hypotheses are proposed.

In contrast with these comprehensive works, and almost as a consequence of ink-on-paper publication, most of the taxonomic literature is composed of short contributions in which one or a few species or genera are described, all too often with minimal taxonomic context. Thus a very large body of basic biodiversity information exists in a highly fragmented state. Users of taxonomic information are faced, therefore, with a disparate information base of many short papers and, with luck, a few consolidated treatments, most of which are likely to be considerably outdated. The last global taxonomic monograph of even as conspicuous a group of insects as the hawkmoths (Lepidoptera: Sphingidae) was published over 100 years ago (Rothschild & Jordan 1903). Much excellent (and considerably less than excellent) work on this family of moths has been published since that time (including the description of numerous species and genera and a full checklist, Kitching & Cadiou 2000): but there is no single up-to-date compilation on this important group of invertebrate animals.

The medium of the Internet offers a solution to this problem, which is why a growing body of literature champions its use in descriptive taxonomy (e.g. Godfray 2002; Scoble 2004; Wheeler 2004; Godfray *et al.* 2007; Scoble *et al.* 2007). The attractive possibility arises of enabling the taxonomy of a group of organisms to be consolidated at a single site on the Web, rendered freely accessible to all users with a connection, and capable of being updated incrementally with new knowledge. A further advantage is that the amount of space for posting text, illustrations, video, audio and alternative taxonomic hypotheses is unlikely to be a limiting factor. Apart from restrictions on space and the capacity to handle different media formats such as video, the prevailing ink-on-paper system means that consolidated works are patchy, up-to-date

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

versions are rare (since they become outdated as soon as, or before, they are printed), accessibility is often restricted to those with access to comprehensive libraries and new information gets published mostly as one-off papers independent of the consolidated monograph. Ink-on-paper survives, however, because its permanence renders it an attractive medium to a discipline as dependent on legacy literature as taxonomy. The legacy archive in taxonomy is crucial for it allows taxonomists and others to locate not simply factual information about a species, but also the taxonomic sense (concept) applied to that species by an author—that is, how did the author delimit the species in terms, say, of the specimens examined, its distribution and the set of characters on which it was defined? Botanists use the word 'protologue' for the published outcome of this holistic exercise.

The limitations of ink-on-paper combined with the existence of the Internet means that Web-based taxonomy has become inevitable. The need for access to biodiversity information in collections of natural history specimens and natural history libraries (particularly their inherent time dimension) provides the major driver to digitize the content in individual institutions and in networking the resultant distributed databases. Already large quantities of infrastructural information are being posted as they become digitized (e.g. on specimens and collection-level metadata (see Scoble & Berendsohn 2007; www.biocase.org)), species-level data (Species 2000/Catalogue of Life http://www.sp2000.org/) and biodiversity literature (Biodiversity Heritage Library, http://www.biodiversitylibrary.org/, a module of the Encyclopedia of Life initiative, www.eol.org). These are large, organized projects, but much taxonomic information about species and higher taxa is being posted on webpages by individuals or special interest groups where little quality control exists or where the level of quality control is not explicit.

We consider that this lack of quality control, the failure to make it explicit or both pose a major problem for Web-based taxonomy. When compiling dictionaries, a wide search may be made for the meanings of words, definitions, derivations and authoritative content, but balanced synthesis and tight editorial control of the final product are essential if such works are to provide the authority rightly expected of them. The analogy with descriptive taxonomy is close. At the same time, Web access to information has become not just an expectation but also an imperative in science as in most areas of human endeavour. So the challenge for taxonomy is how to meet this imperative while providing quality control in a fundamentally anarchic medium. If taxonomists wish to maintain their position as the primary providers of high-quality data on biodiversity, their websites must attract the Web traffic of diverse users to retain their status as the primary brokers of basic information on the fauna and flora of the world. Users already have access to other Web-based sources: they will have many more in the near future.

The CATE project is an attempt to capitalize on the benefits of the Internet as the primary medium for taxonomy and to address the challenges of moving descriptive taxonomy from ink-on-paper to the Web. The project is by no means the only Web-based taxonomic initiative, but we believe its core features are unique in combination. The CATE model seeks to post on the Web a consolidated taxonomy for a given group of organisms, with the capacity to allow the community to add new knowledge online in the form of proposals—such as new species or synonymies. The workflow has been designed to incorporate an online

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

peer-review system allowing editors to add proposals that are accepted, after review, to the consolidated taxonomy and place those not accepted for the consolidated taxonomy on a separate, but accessible, part of the website so that they are not lost. Updatability is a key feature of the system, the aim being to allow new taxonomic information to be added incrementally as it is discovered. Further details of the system and its architecture are described below. We envisage CATE content to be built, reviewed and maintained by taxonomists. As such, the project contrasts with efforts to use aggregation technology to scrape scattered data from the Web (see Butler 2006)—even if those building CATE sites will use information from aggregations as a source on which to apply their taxonomic judgement. But CATE, at heart, relies on traditional taxonomic skills—the kind of skills that all the novel techniques in the world cannot replace.

2. System and methods

(*a*) *Aims*

The CATE application is designed to mount multiple taxonomic hypotheses and to provide a dynamic consolidated taxonomy for a single major group of organisms (such as a family). It aims to be useful to anyone who requires authoritative information on the classification and biology of a group of organisms, whether professional biologists or interested amateurs.

CATE is able to serve taxonomic and nomenclatural data following the traditions, rules and protocols of zoological and botanical taxonomy, respecting the codes of nomenclature of both (ICZN 1999; McNeill *et al.* 2006). Software was developed to support two demonstrator family-level Web taxonomies, one governed by the zoological code of nomenclature and the other by its botanical counterpart (figure 1). Initially, it was expected that these divergent requirements would be challenging to accommodate within a single application. It has proved possible, however, to use a single data model to support both Web revisions particularly because the majority of the customization required by the two traditions can be deferred until the data are rendered as webpages. Differences between the nomenclatural codes and taxonomic traditions affect the presentation of data, but do not prevent the development of generic applications dealing with both. Each code has differing traditions (for example, zoologists cite only the author of the original species name whereas botanists also cite the author, if different, who placed the species into its current genus). In addition, there are different working practices for particular taxa. In our case, aroid taxonomists place great reliance on dichotomous keys, but these are seldom used by sphingid workers who instead favour illustrated diagnoses.

The development of the application began with a prolonged period of requirements gathering to elicit use cases describing how CATE was intended to be used, with the taxonomists developing the content for the two exemplar Web revisions serving as representative users of the application. The application was developed in a series of time-boxed iterations, with priority features and detailed requirements being specified by the users at the beginning of an iteration. At the end of the iteration, the software was deployed, and new content uploaded, allowing the users to validate their data and test the new features of the software

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009



Figure 1. Examples of organisms belonging to the two Web revisions produced by the CATE project. (*a*) One is a global treatment of the hawkmoths (Sphingidae; http://www.cate-sphingidae.org). The hawkmoth is a female of *Deilephila elpenor elpenor* (Linnaeus 1758), from the Czech Republic, taken by Jean Haxaire. (*b*) The other is a global treatment of the Araceae (http://www.cate-araceae.org), a family of flowering plants that include the arum lilies. The plant is *Monstera adansonii* Schott, taken by Ivanilza M. Andrade, in Ceará, Brazil.

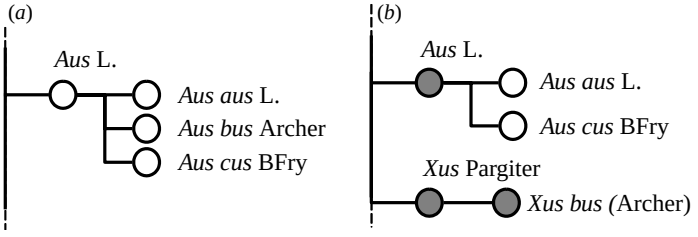


Figure 2. A specific example of a taxonomic change to a Web revision. (*a*) *Sensu* cate-aidae.org (2008): an imaginary revision of the family Aidae contains a genus *Aus* L. with three child taxa *aus* L., *bus* Archer and *cus* BFry. (*b*) *Sensu* Pargiter (2008) proposes that the existing classification be amended, specifically by creating a new genus *Xus* Pargiter, and by amending *Aus* L., transferring *bus* Archer into *Xus*. Circles represent taxon concepts, filled circles (in (*b*)) represent taxon concepts that are introduced or changed in Pargiter's proposal.

prior to a release. The constant focus on users, their requirements and testing using real datasets aimed to ensure that the application met genuine user needs (Neale *et al.* 2007).

To allow the Web resource to remain up-to-date, the system when fully developed will provide tools that allow taxonomists to communicate and propose taxonomic changes to the consensus classification, such as the addition of a description of a new species or the synonymy of the names of two existing taxa. To ensure that the consensus classification reflects the best estimate of the community that has contributed to it, the Web application supports an open peer-review-based workflow that enables contributors to propose new taxonomic hypotheses where they believe the current classification can be improved. To illustrate this process in more concrete terms, consider a CATE site for a hypothetical group Aidae (following the history of the imaginary genus *Aus*, fig. 1, Kennedy *et al.* (2005); see figure 2).

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

(i) The current consensus classification of the group is published online. At this point, the consensus is that *Aus* L. contains three child taxa, *aus* L., *bus* Archer and *cus* BFry.

(ii) Pargiter creates a new proposal entitled 'New molecular phylogeny supports polyphyletic status of *Aus* L. *sensu* cate-aidae.org, 2008'. They propose that the genus be split and that *bus* Archer be assigned to a new genus *Xus*. They create a new genus page for *Xus*, and insert a diagnostic description. They specify that *bus* Archer is the type species of *Xus*. They add additional content to the genus page, and amend the species page of *bus* Archer to correctly refer to *Xus* as its genus. Depending upon the taxonomic code, *Aus bus* Archer might be displayed as a nomenclatural (objective) synonym of *Xus bus* (Archer). In addition to the proposed changed taxon pages, Pargiter fills in a short abstract summarizing the proposed changes, and references the work that led to the proposal.

(iii) Pargiter submits the proposal, and the proposal is advertised in various ways: on the affected pages (*A. bus* Archer and the *Aus* L. generic page), via email alerts to users who have subscribed to them and via Web feeds. Users are able to visit the proposal page and browse the proposed changed taxon pages.

The Web-based nature of the revision allows the quality of the proposal to be assessed using multiple methods, including permitting registered users to suggest improvements or corrections, 'voting' by users that a proposal be accepted or rejected, soliciting reviews from specialists by the editorial board or algorithmic means of assessing proposals (e.g. checking that proposals contain mandatory information). Depending upon the taxonomic proposal and the community that uses the Web revision, proposals may be treated with different levels of consideration. In the case of small communities, non-controversial proposals and small changes or corrections to the classification, changes may be accepted and incorporated into the revision without further debate. In the case of more controversial proposals, the taxon committee will be required to decide when a proposal has been reviewed to a sufficiently high standard that a decision can be made.

(iv) In this case, the proposal that *Aus* L. be split is controversial. Several members of the community recommend that the proposal is rejected and cite their reasoning. On balance, the committee decides that the proposal is not supported and that it should be hosted on the site as an alternative hypothesis. Users are able to access the classification as proposed by Pargiter, and the proposal itself. In the published consensus classification, the genus *Xus* Pargiter is regarded as a new synonym of *Aus* L. *sensu* cate-aidae.org, and *X. bus* (Archer) is regarded as a new synonym of *A. bus* Archer.

Importantly, changes to the consensus classification are mediated by taxonomists: while new proposals are automatically mounted on the site, their incorporation into the consensus taxonomy requires peer review and a decision by an editorial committee for that taxon. All comments by reviewers and editors will be made available on the website, alongside alternative taxonomic hypotheses and previous versions of the consensus classification. The process of updating is thus fully transparent and traceable.

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

Where a taxonomic change proves to be very controversial (as in the case of the genus *Acacia* Mill. (Smith *et al.* (2006)), or the revised classification of the Aedini (Diptera: Culicidae) (Polaszek 2006)), it seems likely that wider debate and decisions taken by other organizations (such as the commissions responsible for the codes of nomenclature) would affect the final classification. In situations such as these, the taxon committee can take the decision provisionally to accept or reject a proposal, amending the consensus classification if necessary once the outcome is clear. The advantage of a Web-based revision in such situations is threefold: firstly that areas of uncertainty or conflict can be highlighted to users and linked to further information; secondly that publishing a revised classification can be accomplished with much greater ease than in the case of paper publication; finally that users do not need to wait until the debate has been resolved—globally unique identifiers (GUIDs) allow end-users to identify which taxonomic concept they are using in a manner that allows data to be reinterpreted if the classification changes later.

The taxon committee plays a role similar to the editorial board of a journal, being responsible for both the content and the quality of the Web taxonomy and the practicalities of running the site and liaising with the organization hosting the software. Were Web taxonomies to become the norm, it would be important to create mechanisms to ensure taxon committees were truly representative of the communities they serve. Overseeing this process might be a role taken on by the organizations that currently administer the nomenclatural codes. In the case of the two CATE exemplars, the editorial committees have been drawn from the teams working on the project and/or experts external to the teams.

The consensus or consolidated classification is the sum of numerous individual contributions made by many authors, past and current, and it is critical that each contribution is recognized. At the same time, it is important that the classification is freely available to be used as the basis for further scientific research. The authorship of information ('content') contributed to a CATE Web revision is associated with each item of data (e.g. name, specimen, description, reference, image). In addition, summary statistics for contributions by a particular author can be produced, providing evidence of the author's overall contribution to the Web revision. The content of the Web revision is made available under an open-access licence such as the creative commons attribution/non-commercial/share-alike licence (http://creativecommons.org).

(*b*) *Design and workflow*

Data are primarily accessed using a Web browser and displayed in a format reminiscent of a paper monograph (figure 3). The data presented in a taxonomic revision are diverse and consequently the data model is complex. The data model mostly concerns taxonomic and nomenclatural information, and descriptive data. Other types of information covered include publications, people, organizations, specimens and observations, images, molecular and geographical data—those data expected in modern taxonomic works.

CATE takes advantage of its complex data model, and third-party software to enhance access to the data it contains, allowing users to interact with content. Textual content (e.g. a taxonomic name) within the application is indexed and can be searched. Information about species can also be found by browsing the taxonomic hierarchy or by searching for taxa found in a specific geographical

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

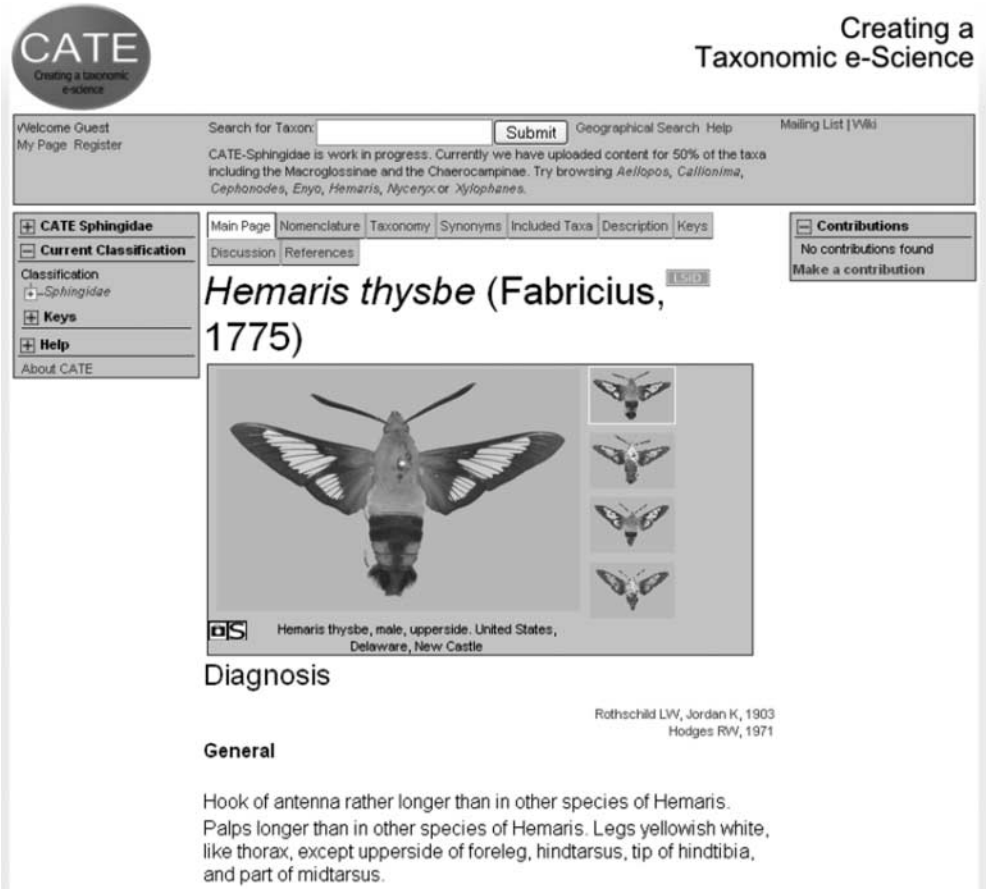


Figure 3. A screenshot of the current release of the CATE Sphingidae exemplar site, opened on the taxon page for *Hemaris thysbe* (http://www.cate-sphingidae.org).

area. Images are provided in a format that can be zoomed and panned, allowing users to view them in great detail. Keys are constructed using the Lucid Player Java applet (http://www.lucidcentral.org) that presents them in an interactive, multi-access form with illustrations and definitions of characters and states.

CATE goes beyond a static resource, and provides the means by which taxonomists can continue to revise and update the taxonomy of the group in question. To allow the evolving classification to be tracked in time, changes do not lead to data being overwritten. Instead, CATE follows an incremental cycle of revision and publication, with the current consensus classification and alternative hypotheses being presented to end-users, but with earlier versions and withdrawn hypotheses being preserved and archived. As described earlier, new proposals are refereed and opinions sought from the whole community before the taxon committee decides whether they should be incorporated into the next edition of the consensus taxonomy.

Proposing new taxonomic hypotheses, their review and possible incorporation in the next edition consensus taxonomy requires communication among the taxonomists working on the same group. Rather than develop collaborative

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

software de novo (for example, to allow mailing lists, forums, file-sharing, news feeds and wiki-like functionality), a customized instance of the popular Content Management System *Drupal* (http://drupal.org/) is used to provide this functionality to the community of users. The resultant 'scratchpads', produced by the European Distributed Institute of Taxonomy (EDIT; Work Package 6 Scratchpad project, http://www.editwebrevisions.info/scratchpads), provide tools for the community of taxonomists to share files, collaboratively edit documents and communicate with each other, facilitating taxonomic research. The two software systems share the same URL, residing at, for example, www.cate-araceae.org and scratchpad.cate-araceae.org, but are not yet integrated to the degree that data are shared between them.

3. System architecture

The CATE Web application is implemented in Java, using popular open-source software (MySQL, a database server: http://www.mysql.com; Spring Framework, a Java application framework: http://www.springframework.org; Hibernate, an object-relational persistence and query service: http://www.hibernate.org; and Apache Tomcat, a Java servlet container: http://tomcat.apache.org) and is itself available under an open-source licence. It is based upon the Common Data Model (CDM) Server application produced by EDIT (Work Package 5 Internet Platform for Cybertaxonomy, http://wp5.e-taxonomy.eu/). The EDIT CDM is a data model for biodiversity and particularly taxonomic data, and is in turn based upon existing biodiversity informatics standards (particularly the ontology developed by Biodiversity Information Standards organization, http://www.tdwg.org).

The CDM server is a software application that provides more generic services such as persistence and retrieval, query and remote access to data stored on the server using a number of standard protocols. CATE augments the CDM library and CDM server with additional logic to present the data as webpages and support the workflow of revisionary taxonomy (figure 4).

A CATE Web taxonomy is designed to be a high-quality, up-to-date, taxonomic resource. Seamless and open integration into the global biodiversity informatics landscape allows the full benefit of the taxonomic effort expended in creating and maintaining a site to be realized (Page 2006; Patterson *et al.* 2006; Guralnick *et al.* 2007; Pendry *et al.* 2007). In particular, providing the consensus classification as a Web service allows other applications to use the most up-to-date consensus classification as the basis for further analysis. CATE achieves this by assigning Life Science Identifiers (LSIDs) to the taxonomic concepts in its classification (Clark *et al.* 2004; Object Management Group 2004). LSIDs provide a standard way of assigning and resolving GUIDs for life sciences data. GUIDs are digital identifiers that can be used to identify objects or concepts uniquely and to retrieve metadata about them (Digital Object Identifiers, http://www.doi.org/, are an example of another GUID technology used to identify digital documents). LSIDs also allow CATE to link to other biodiversity resources by returning references to non-CATE LSIDs in metadata. For example, metadata about the taxonomic concepts in cate-araceae.org include references to the LSID for the taxonomic name of that concept served by the International Plant Names Index (Croft *et al.* 1999) where the identifier is available. The ability to assign identifiers

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

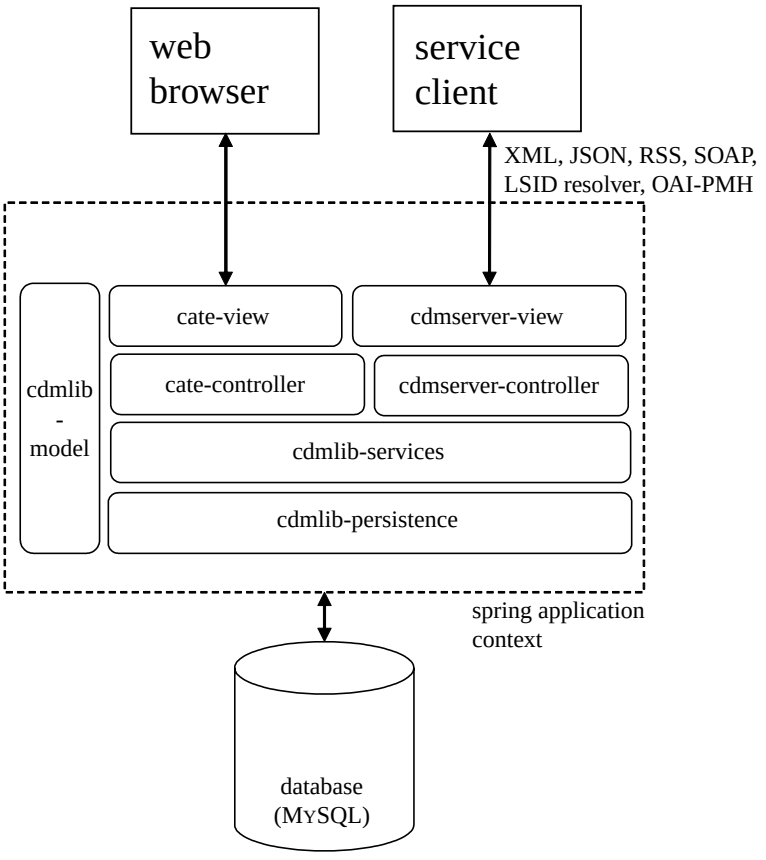


Figure 4. A high-level diagram summarizing the architecture of the CATE application. Each box represents a software component. CATE is built upon the EDIT CDM and Java Library; CDM objects are exchanged between layers of the application in response to a request from a client, such as a Web browser. The lowest layer, cdmlib-persistence, provides generic access to the database and the next layer, cdmlib-services, provides an application programming interface to the upper layers of the application. It is shared by both CATE and other software based upon the CDM. The CDM server exposes data as Web services (using the cdmserver-controller and cdmserver-view components); these allow other applications to access the data in the Web revision directly. In addition, CATE presents the data as a set of webpages, and provides the means for interacting with the data via a Web browser (using the cate-controller and cate-view components).

to specific taxon concepts is important in the context of a dynamic Web revision because it allows software clients to distinguish between the senses in which a taxon name is being used in different treatments and to retrieve the most recent metadata about the taxon concept automatically.

4. Discussion

Three areas of the CATE project have elicited particular questions: is peer review the best model for quality control; does 'consensus' taxonomy not impose too great a restriction; and can community style taxonomy be made sufficiently

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

attractive to its practitioners such that it makes an impact on the taxonomic impediment? We comment, briefly, on the first two of these questions; our thoughts on the third are expressed in our responses to the others.

(*a*) *Peer review*

We have opted to embed a peer-review process in the CATE workflow. Although conscious that alternatives exist (e.g. books are reviewed *after* their publication by Amazon.com and eBay), the traditional model delivers, albeit imperfectly, an explicit process of quality control. Many areas of research benefit from this approach. But taxonomy is an additive discipline, where new taxa and nomenclatural changes remain part of the knowledge base. There is more than enough risk in moving taxonomy to the Web: tried and tested procedures in cleansing the knowledge base have their place.

The open peer-review process proposed in the workflow (see above) is intended to deliver high-quality taxonomic information without the process becoming an impediment. The number of submissions and reviews made, and the amount of effort expended by users of the site, will depend upon a Web revision being sufficiently attractive to motivate and recruit users. Studies of other online projects (e.g. Lakhani & Wolf 2003) suggest that users of Web revisions may be expected to contribute for a variety of reasons. For most professional taxonomists, contributing effort to enhance online resources will be compelling as long as their efforts are recognized and encouraged by their respective institutions (Knapp 2008). Attributing credit to those who create or curate data in large projects, and providing measures of the value of contributions, is necessary to reward individuals for the work done (Birnholtz 2006).

(*b*) *Consensus taxonomies*

The concept of dynamic Web taxonomies (Godfray 2002), especially the provision of a consensus classification, has led to concern that the complex process of revisionary taxonomy might become oversimplified, giving the impression that there is 'one true taxonomy' (Thiele & Yeates 2002). Moreover, it has been argued that a comprehensive and reliable classification can be the product of only thorough research and debate by experts, and that replacing the complexities of the current paper-based taxonomy with what is perceived as simplified Web-based taxonomy is a retrograde step (de Carvalho *et al.* 2007). These concerns seem to have arisen through misunderstanding. First, we suggest that consensus taxonomies already exist: major ink-on-paper taxonomic revisions often acquire this status de facto (Scoble *et al.* 2007). Collaborative work has produced large 'consensus' floras. The demand for synthetic works on a greater range of organisms testifies to the need of consumers of taxonomy for comprehensive, authoritative and rapidly updatable treatments. The CATE workflow aims to capture the dynamic process of revisionary taxonomy and enable taxonomists to meet the need (e.g. for conservation, see Mace 2004) for up-to-date taxonomic classifications.

Second, a CATE-style consensus taxonomy is intended to retain alternative views on the website (as described earlier), so that they can, potentially, be revived. Good revisionary taxonomy, whether Web or paper based, explains

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

differences of opinion but still proposes a recommendation. Consensus, therefore, is neither intended to stifle dissent nor does it imply immutability. It is needed to help users outside the taxonomic community, who neither wish to choose between competing views (at least at the same source) nor, probably, are equipped to do so. Why, to reverse the question, should users *not* expect taxonomists to provide their best view at a given time? We accept that achieving a consensus will not always be easy (Vane-Wright 2003), and that sometimes it may prove impossible, but that should not stop us trying.

Moving taxonomy to the Web as a truly open, collaborative and dynamically updatable exercise presents challenges as well as opportunities for individuals and institutions (Hine 2008), requiring communication and coordination to achieve a shared goal. CATE and other software applications can facilitate this process, but ultimately it will be the communities of taxonomists that will make it a success: forging successful online communities is just as demanding as creating the software that supports them.

Coordination of effort across a distributed community of taxonomists, managing intellectual property and rewarding those who contribute effectively (David 2004; Burk 2007) are challenges that need to be addressed if CATE or related projects are to have any hope of success. As with other online collaborative projects, much of the infrastructure developed is aimed at facilitating communication and collaboration.

The authors thank two referees for their valuable comments and express their gratitude to the Natural Environment Research Council for providing support under grants NE/C001532, NE/C51588X/2 and NE/C515871/1.

References

Birnholtz, J. P. 2006 What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *J. Am. Soc. Inf. Sci. Technol.* **57**, 1758–1770. (doi:10.1002/asi.20380)

Burk, D. L. 2007 Intellectual property in the context of e-Science. *J. Comput. Med. Commun.* **12**, 600–617. (doi:10.1111/j.1083-6101.2007.00340.x)

Butler, D. 2006 Mashups mix data into global service. *Nature* **439**, 6–7. (doi:10.1038/439006a)

Clark, T., Martin, S. & Liefeld, T. 2004 Globally distributed object identification for biological knowledgebases. *Brief. Bioinform.* **5**, 59–71. (doi:10.1093/bib/5.1.59)

Croft, J., Cross, N., Hinchcliffe, S., Nic Lughadha, E., Stevens, P. F., West, J. G. & Whitbread, G. 1999 Plant names for the 21st century: the International Plant Names Index, a distributed data source of general accessibility. *Taxon* **48**, 317–324. (doi:10.2307/1224436)

David, P. A. 2004 Towards a *cyberinfrastructure* for enhanced scientific collaboration: providing its 'soft' foundations may be the hardest part. SIEPR discussion paper no. 04-01. Stanford Institute for Economic Policy Research, Stanford University, Stanford, CA.

de Carvalho, M. R. *et al.* 2007 Taxonomic impediment or impediment to taxonomy? A commentary on systematics and the cybertaxonomic–automation paradigm. *Evol. Biol.* **34**, 140–143. (doi:10.1007/s11692-007-9011-6)

Godfray, H. C. J. 2002 Challenges for taxonomy. *Nature* **417**, 17–19. (doi:10.1038/417017a)

Godfray, H. C. J. & Knapp, S. 2004 Taxonomy for the twenty-first century. *Phil. Trans. R. Soc. B* **359**, 559–739. (doi:10.1098/rstb.2003.1457)

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

Godfray, H. C. J., Clark, B. R., Kitching, I. J., Mayo, S. J. & Scoble, M. J. 2007 The Web and the structure of taxonomy. *Syst. Biol.* **56**, 943–955. (doi:10.1080/10635150701777521)

Guralnick, R. P., Hill, A. W. & Lane, M. 2007 Towards a collaborative, global infrastructure for biodiversity assessment. *Ecol. Lett.* **10**, 663–672. (doi:10.1111/j.1461-0248.2007.01063.x)

Hine, C. 2008 *Systematics as cyberscience: computers, change, and continuity in science*. Cambridge, MA: MIT Press.

ICZN 1999 *International code on zoological nomenclature*. London, UK: International Trust for Zoological Nomenclature.

Kennedy, J. B., Kulka, R. & Patterson, T. 2005 Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In *Data Integration in the Life Sciences: 2nd International Workshop, DILS 2005, San Diego, CA, 20–22 July 2005*. Lecture Notes in Computer Science, no. 3615, pp. 80–95. Berlin, Germany: Springer.

Kitching, I. J. & Cadiou, J.-M. 2000 *Hawkmoths of the world: an annotated and illustrated revisionary checklist*. London, UK: The Natural History Museum; New York, NY: Cornell University Press.

Knapp, S. 2008 Taxonomy as a team sport. In *The new taxonomy* (ed. Q. D. Wheeler), pp 33–54. Boca Raton, FL: CRC Press/Taylor & Francis Group.

Lakhani, K. R. & Wolf, R. G. 2003 *Why hackers do what they do: understanding motivation and effort in free/open source software projects*. MIT Sloan School of Management, Working Paper Series, no. 4425-03.

Linnaeus, C. 1753 *Species plantarum*, 1st edn. Holmiae (Stockholm), Sweden: Salvius.

Linnaeus, C. 1758 *Systema naturae*, 10th edn. Holmiae (Stockholm), Sweden: Salvius.

Mace, G. 2004 The role of taxonomy in species conservation. *Phil. Trans. R. Soc. B* **359**, 711–719. (doi:10.1098/rstb.2003.1454)

McNeill, J. *et al.* 2006 *International code of botanical nomenclature* (*Vienna Code*). Ruggell, Liechtenstein: A. R. G. Gantner Verlag.

Minelli, A. 2003 The status of taxonomic literature. *Trends Ecol. Evol.* **18**, 75–76. (doi:10.1016/S0169-5347(02)00051-4)

Neale, S. H., Pullan, M. R. & Watson, M. F. 2007 Online biodiversity resources—principles for usability. *Biodivers. Inform.* **4**, 27–36.

Object Management Group 2004 *Life sciences identifiers specification*. See http://www.omg.org/cgi-bin/doc?dtc/04-05-01.

Page, R. D. M. 2006 Taxonomic names, metadata, and the Semantic Web. *Biodivers. Inform.* **3**, 1–15.

Patterson, D. J., Remsen, D., Marino, W. A. & Norton, C. 2006 Taxonomic indexing—extending the role of taxonomy. *Syst. Biol.* **55**, 367–373. (doi:10.1080/10635150500541680)

Pendry, C. A., Dick, J., Pullan, M. R., Knees, S. G., Miller, A. G., Neale, S. & Watson, M. F. 2007 In search of functional flora—towards a greater integration of ecology and taxonomy. *Plant Ecol.* **192**, 161–167. (doi:10.1007/s11258-007-9304-y)

Polaszek, A. 2006 Two words colliding: resistance to changes in the scientific names of animals—*Aedes* vs *Stegomyia*. *Trends Parasitol.* **22**, 8–9. (doi:10.1016/j.pt.2005.11.003)

Rothschild, W. & Jordan, K. 1903 A revision of the lepidopterous family Sphingidae. *Novit. Zool.* **9**, 1–972.

Scoble, M. J. 2004 Unitary or unified taxonomy? *Phil. Trans. R. Soc. B* **359**, 699–710. (doi:10.1098/rstb.2003.1456)

Scoble, M. J. & Berendsohn, W. 2007 Networking biological collections databases. In *Biodiversity databases: techniques, politics, and applications* (eds G. B. Curry & C. H. Humphries), pp. 23–35. Boca Raton, FL; London, UK; New York, NY: CRC Press.

Scoble, M. J., Clark, B. J., Godfray, H. C. J., Kitching, I. J. & Mayo, S. J. 2007 Revisionary taxonomy in a changing e-Landscape. *Tijdschr. Entomol.* **150**, 305–317.

Smith, G. F., van Wyk, A. E., Luckow, M. & Schrire, B. 2006 Conserving *Acacia* Mill. with a conserved type. What happened at Vienna? *Taxon* **55**, 223–225.

Downloaded from rsta.royalsocietypublishing.org on 16 June 2009

Stork, N. E. 1993 How many species are there? *Biodivers. Conserv.* **2**, 215–232. (doi:10.1007/BF00056669)

Thiele, K. & Yeates, D. 2002 Tension arises from the duality at the heart of taxonomy. *Nature* **419**, 337. (doi:10.1038/419337a)

Vane-Wright, R. I. 2003 Indifferent philosophy versus almighty authority: on consistency, consensus, and unitary taxonomy. *Syst. Biodivers.* **1**, 3–11. (doi:10.1017/S1477200003001063)

Wheeler, Q. D. 2004 Taxonomic triage and the poverty of phylogeny? *Phil. Trans. R. Soc. B* **359**, 571–583. (doi:10.1098/rstb.2003.1452)

Wheeler, Q. D., Raven, P. H. & Wilson, E. O. 2004 Taxonomy: impediment or expedient? *Science* **303**, 285. (doi:10.1126/science.303.5656.285)