Acta Protozool. (2014) 53: 115–128
http://www.eko.uj.edu.pl/ap
doi:10.4467/16890027AP.14.011.1448

ACTA
PROTOZOOLOGICA

Special issue:
Marine Heterotrophic Protists
Guest editors: John R. Dolan and David J. S. Montagnes

Review paper

Helping Protists to Find Their Place in a Big Data World

David J. Patterson

School of Life Sciences, Arizona State University, Tempe, Arizona, USA

Abstract. The 'big new biology' is a vision of a discipline transformed by a commitment to sharing data and with investigative practices that call on very large open pools of freely accessible data. As this datacentric world matures, biologists will be better able to manage the deluge of data arising from digitization programs, governmental mandates for data sharing, and increasing instrumentation of science. The big new biology will create new opportunities for research and will enable scientists to answer questions that require access to data on a scale not previously possible. Informatics will become the new genomics, and those not participating will become marginalized. If a traditional discipline like protistology is to benefit from this big data world, it must define, build, and populate an appropriate infrastructure. The infrastructure is likely to be modular, with modules focusing on needs within defined subject and makes it available in standard formats by an array of pathways. It is the responsibility of protistologists to build such nodes for their own discipline.

Key words: Big data, Biodiversity informatics, protistology, name-based cyberinfrastructure.

A BIG DATA WORLD

More of the sciences are expected to metamorphose around data sharing and data re-use (NSF 2006, 2011, National Research Council of the National Academies 2009; Hey *et al.* 2009; Wood *et al.* 2010). While the vision of the emerging 'Big Data' world was the subject of special issues of Nature (Big Data special, volume 455, 2008), and Science (Dealing with big data, volume 331, 2011), the topic barely figures within protistology.

There are three primary reasons why we should invest in a transformed discipline. All are applicable to protistology. The first is to intercept, manage, and exploit the data deluge that results from increased instrumentation, environmental monitoring, digitization programs, collapsing sequencing costs, and mandates from funding agencies (e.g. Baker 2010, Kelling *et al.* 2009, Wetterstrand 2013, Wood *et al.* 2010). Secondly, the internet has become a virtual data pool that has created new opportunities for discovery (Gore 2013). Finally, scientists want to address problems that require more information than can be created by small research teams (Caron and Hutchins 2013, Sarmento *et al.* 2010). Additional motivators include economic efficiencies of re-using rather than recreating data (Piwowar *et al.* 2011), governmental pressures for open-ness (URL 1), or making irreplaceable legacy data, such as where radiolaria occurred in the 19th century (Haeckel 1887), available for general re-use.

Address for correspondence: David J. Patterson, School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501, USA; E-mail: david.j.patterson@asu.edu

A 'Big New Biology' is envisaged as a discipline with a strong data-centric character and a growing role for informatics (National Research Council of the National Academies 2009, Patterson 2009, Patterson *et al.* 2010). If we wish to position protistology in this emerging world, we will need to set priorities, design, build, organize, and populate the infrastructure that will serve our research agendas; and become better motivated to share content (Thessen and Patterson 2011). Change will bring research and career opportunities.

FIRST, WE NEED TO MAKE CONTENT AVAILABLE

A pre-requisite for participation in the big data world is that data are rendered digital and are made available on-line. That we still have some way to go is illustrated with on-line information about the distribution of a tintinnid, *Rhabdonellopsis apophysata*. A literature-based survey (Pierce and Turner 1993) is our comparator. The Global Biodiversity Information Facility (GBIF, URL 2) and the Ocean Biogeographic Information System (OBIS, http://www.iobis.org/ URL 3) collect data on georeferenced occurrences of species but have no information on *R. apophysata*. Sequence databases such as GenBank (URL 4) may refer to where samples came from, but not for this tintinnid. Pangaea (URL 5) as a major data repository for Earth system research has 20 files that include data for this species. This helps, but the absence of these data from GBIF and OBIS reveals that data are not yet flowing from one location to another. The older literature is increasingly available on-line in the Biodiversity Heritage Library (URL 6), but a search produces a single result (Pierce and Turner used information from 272 sources). New data may appear in web sites such as the World Register of Marine Species – which tells us that R. *apophysata* occurs in the Gulf of Mexico (URL 7) whereas Pierce and Turner provide about 100 datum points that show that the species has a circum-global distribution. Other sites that may contain data include Tree of Life (URL 8), micro*scope (URL 9), the Plankton Ciliate project (URL 10), the Villefranche sur mer web site (URL 11), the Protist Information Server (URL 12), the Checklist of Phytoplankton in the Skagerrak-Kattegat (URL 13), and so on. Despite the promise of aggregating initiatives such as the Encyclopedia of Life (URL 14), DiscoverLife (URL 15), or Atlas of Living Australia (URL 16), they offer little of relevance, the former picking up only the misleading 'Gulf of Mexico' information from WoRMS.

Turning to search engines, Google finds about 60 sites with information on the species, about half providing distributional data. Pierce and Turner's map of *R. apophysata* has about 100 points. Taxonomic changes can make data hard to find. A search needs to know all of the names that have been used for a taxon; in this case, that *R. apophysata* was introduced as *Cyttarocylis hebe* var. *apophysata* and has been known as *Cyttarocylis apophysata*. The addition of synonyms adds 20 more useful pages to Google's find. The addition of such expert knowledge into general functions is referred to as 'taxonomic intelligence.' Synonymy information is not easy to find. The Catalogue of Life (URL 17) has neither senior synonym nor junior synonyms for this species. The junior synonym is not available on WoRMS, but is available on the Marine Species Identification portal (URL 18). Again, the isolation of the information shows that content is not flowing as is envisaged by the big data vision.

This poor representation of biodiversity information on the internet is not unusual. Thessen *et al.* (2012) found that only about 30% of the information that related to the taxonomy of *Gymnodinium* was available on-line. The internet is still not the place to go to for expert protistological information. We can change that. There are free generic environments such as ScratchPads (URL 19) that make participation simple and free.

WHAT MIGHT AN INFRASTRUCTURE LOOK LIKE?

Given what has emerged already, the responsibility for managing data from many sources will probably be carried out by modules that serve specified subdisciplines. Modules will have sources and serve users with nodes taking responsibility for delivering content in consistent formats (Fig. 1a). The role of the sources is to make content available. Users take responsibility for visualization, analysis and synthesis of shared data; and can play an important role in quality control (see section on Annotation). Nodes will interact (Fig. 1b), adapting and evolving as needs and opportunities arise. The International Nucleotide Sequence Database Collaboration (URL 20) that includes GenBank is our best model of what modules might look like. GenBank pro-

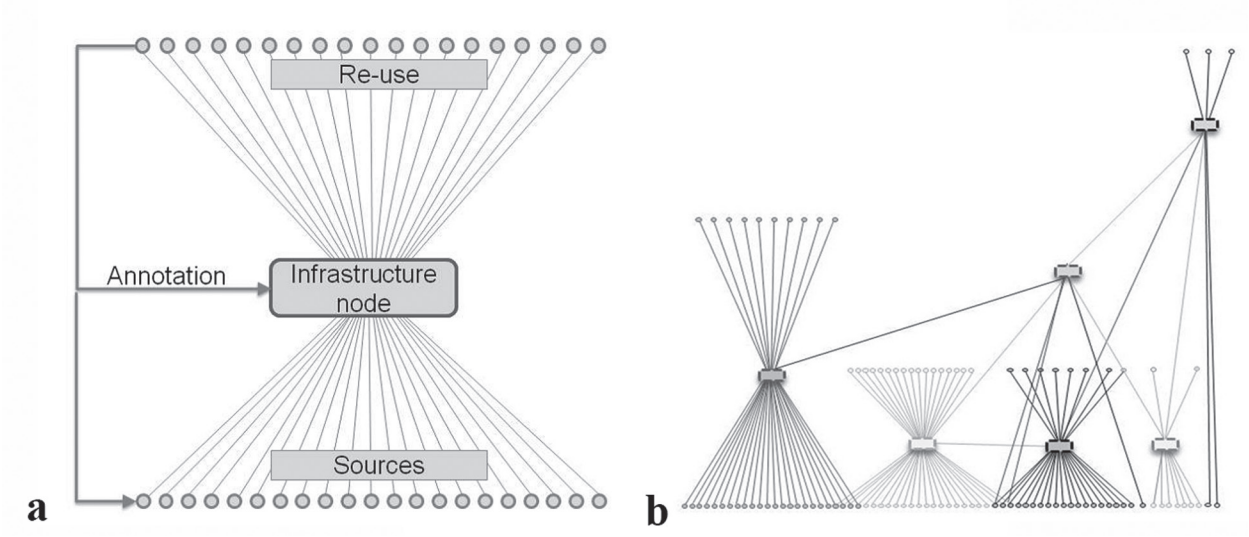

Fig. 1. Modular' model for the infrastructure of a big data world. **a** – Within a module, nodes obtain content from one or more sources, normalize, enrich, and deliver it to end users. Annotation systems allow users to advise the source and nodes as to the quality of content. **b** – Nodes interconnect in anarchic ways that allow for evolution and expanding functionality.

vides access to information from thousands of sources, and is so useful that an array of associated services and derivative products have grown up around it. Other initiatives, such as Catalogue of Life (URL 17) and GBIF (URL 2) have a similar form but are less well developed. There are some initiatives that collate protist data – including the Protist Ribosomal Database (Guillou *et al.* 2013), micro*scope (URL 9), PlasmoDB (URL 21), ToxoDB (URL 22), the World Foraminifera Database (URL 23), and AlgaeBase (URL 24) but have yet to fully participate in active data sharing.

ROLES FOR NODES

The nodes will be most important elements of a future infrastructure. They will aggregate heterogeneous content within a particular subdomain, making it discoverable and available to end users. Some roles of nodes are listed in Box A.

As data are acquired, different systems of units (°F vs °C) need to be transformed into common formats. This is referred to as normalization. In order to foster interoperability and re-use, normalization needs to comply with communally agreed discipline data standards such as those emerging from Biodiversity Information Standards (URL 25). Tools are available to help standardize content (URLs 26–28).

Data are often supplied in data files, such as spreadsheets. If end-users are not to be faced with the task of extracting data from a diversity of original data files so that they can study them, nodes will ideally take responsibility for extracting the smallest effective components from the files. This is referred to as atomization. Ideally, each resulting semantically minimal element (a variable and a value for it) should be labelled with a universally unique and persistent identifier or UUID (GBIF 2011).

Nodes will make content widely available through user interfaces that are used by people, by APIs for use by computers to download content, or by exporting discovery metadata and/or content to the Linked Open Data Cloud (URL 29), ideally applying the Open Archives Initiative Protocol for Metadata Harvesting (URL 30) to broaden access.

Given the inherent dirtiness of bio-data, nodes need to address issues of data quality (Chapman 2005). Quality control by expert scrutiny of data on the way into the system will create bottlenecks because there are few people qualified to vet data and there are rarely any rewards for participating in this process. Algorithmic solutions can be used to validate (establish compliance with standards) data and the inclusion of a data curator / data manager within the team will help. There are also crowd-sourcing solutions to this problem (see 'Annotation' below).

Box A. Some responsibilities of an infrastructural node

- Has a defined scope and purpose
- Knows the requirements of users
- Discovers relevant data / information
- Registers sources of data and information
- Works with sources to make content available
- Acquires data / information
- Stores data, provides preservation and curatorial services
- Converts data files into individual units (atoms) of data
- Converts comparable data to the same system of units (normalization)
- Applies discipline standards and ontologies to content
- Applies universally unique dereferenceable identifiers to data elements
- Organizes the content so that users can gain access to content from many sources at the same time
- Enables access to the data through user interfaces, web services, and via the Linked Open Data Cloud.
- Takes responsibility for data validation and data quality
- Provides taxonomic intelligence to deal with synonyms, homonyms, and to allow content browsing and aggregative searches
- Retains and communicates the provenance of content
- Gives authors of content and intermediaries credit and responsibility for their efforts.
- Develops a business model for sustainability

Nodes will incur costs for hardware, to ensure compliance with best practices, to interact with sources and users, and for future expansion. Grant-based support is a poor model to sustain infrastructure (see next section), and nodes will have to come up with an effective business model for their persistence.

BUILDING INFRASTRUCTURE IS NOT THE SAME AS DOING RESEARCH

The research agenda is to promote discovery. To achieve this, funding agencies invest in many research projects that capitalize on the vision and passion of individuals and small groups to benefit from their originality. The research agenda includes considerable redundancy with many teams targeting the same area – because this increases the probability of a good result in this high-risk activity. The character of research changes as new technologies and paradigms appear, and funding preferences change over time. In that sense research is ephemeral. Any given research project may or may not be funded, so is optional.

The agenda of an infrastructure is to serve. The design, construction and maintenance of infrastructure need to be supported by the community that will benefit from it. It must be reliable, but must be able to evolve to meet new needs. Reliability, in part, demands persistence and long term commitment. Combined, this requires a form of collegiality and long term thinking that does not characterize research. Infrastructure needs a funding model unlike that needed for optional, ephemeral and egocentric research.

The assembly of an infrastructure is less about originality than about implementation. Elements of infrastructure will likely emerge in a series of stages, each distinguished by improvements in reliability and services. The most likely start point will be a 'Proof of Concept' environment that is built the context of a research agenda. Its success confirms that the logic and data model are sound and implementable, are not expensive to build nor are they built to serve a diversity of use cases or users.

With appropriate stakeholder input, domain and informatics expertise, and funding; such systems evolve into prototype services that are aimed at a wider community. Prototype services are also likely to emerge within research environments. Prototypes are not designed to provide a robust infrastructure, to work under all circumstances, to deal with edge cases, or handle heavy demand. We can think of these as services that satisfy users at least 80% of the time, but some users are not well served.

The next level of development ensues if a prototype satisfies its community and becomes critical to the research agenda. The next phase, of production services, moves more into the hands of coding experts who refactor the code-base and impose test protocols to improve the performance of services so that they meet expectations at least 95% of the time, but do not meet the highest expectations for robustness and reliability. The more exacting standards make the development of production software significantly more expensive.

The final level of performance is a flawless system that has little or no down time, rarely fails to meet the expectations of users, performs quickly and impeccable under all circumstances, and is always there. Mature infrastructure should aspire to this level of performance. It is costly to build software that is flawless, perhaps costing 100 times more that prototype services.

METADATA AND THE LIKE

Metadata are critical to finding and using data. Metadata are terms that describe information. There are several classes of metadata, including metadata that

deal with the form of files, provenance of files, describe the files or the content they include, and rights relating to files. Discovery metadata have low precision (are coarse-grained), but help users to find content that may be relevant. At the other end of the spectrum, re-use of data ideally requires metadata with sufficiently fine granularity as to point to data elements (atoms).

Some metadata have a disproportionate power to draw content together because they are associated with many data sets. They are the metadata that define location (georeferencing), date and time, and the names of organisms. This information may often be included within data files as data. Because of their widespread use and predictable form, software has been developed to scrutinize the content of data files, find reference to these 'key integrators,' and extract them for use as metadata that point to individual records. An example are the natural language processing tools that identify and extract the names of organisms in sources (Thessen Cui and Mozzherin 2012). Unfortunately, we find that many data files include idiosyncratic or unhelpful terms as names. A recent survey of the Dryad data repository (URL 31) revealed that the following were offered as names: Aa, Ar, Pet1, A marina, Abe_Heli, Apodemia. mor.A13, N_larina_aethra_20018, Apion pensylvaticum: Boheman 1839, Apion pennsylvaticum Boheman, 1839, Gy091_Lv_Bonn_Ger, P.potto_JCKerbis2889, S.sciereus_U53582, C.major, and L._catta. In many such cases, the correct name cannot be determined. This matter can be addressed if tools are built to check taxonomic names as data are being entered.

Metadata can be made much more useful if their relationships are defined by ontologies. Ontologies can be very complex and there are many of them (URLs 32, 33). The development of ontological frameworks is neither complete nor unified. As a community, we will have to develop ontologies for all descriptive terms that have been used in a domain if we wish to take advantage of machine reasoning. Ontologies do not need to be developed in advance, but can be assembled and applied later in the data life cycle.

THE SIGNIFICANCE OF NAMES

Names are associated with almost every statement about a species in the scientific literature and so can be used as metadata to index and organize data about any species. Because of their potential for indexing content, it is inevitable that a names-based cyberinfrastructure will be a part of the Big New Biology (Patterson *et al.* 2008, 2010). To be effective, the names-based infrastructure will need to embrace the dark taxa known from molecular surveys but not yet with conventional names (Caron *et al.* 2009, Charvet *et al.* 2012, Page 2011, Pawlowski *et al.* 2011).

To serve as metadata, names need to be stable in meaning, and for there to be little or no ambiguity as to what they refer to. The codes of nomenclature seek similar goals by ensuring that every species has a single name, and that each name is used for a single taxon. Unfortunately, the goals are not achieved. Each code is limited scope (to animals, to prokaryotes, or to plants, algae, and fungi) and are independent of each other such that the same name can be applied to a plant and to an animal (etc.). These are homonyms. An estimated 15% of genera are homonyms (URL 34).

The term 'name-string' refers to the sequence of alphanumeric characters with spaces and punctuation that is used as a name. Many different name-strings may apply to the same species – hence *Paramecium aurelia*, *P. aurelia*, *Paramaecium aurelia*, *Paramaecium aurelia*, *P. aurelia* OFM, *Parамecium aurelia* Müller, 1773 are a few of the name-strings that have been legitimately used for the same species. The Global Names Index (URL 35) has over 22 million names for the 2.3 million or so of extinct and living species (Chapman 2009, Raup 1991). Variations in name strings is quantitatively the biggest problem that a names based cyberinfrastructure has to overcome if it is to draw together all available information on the same species.

Other problems that cause problems in using names to index content come from the progress in taxonomy that cause species to be moved from one genus to another. As this happens, binomial combinations change, and homotypic synonyms are born. *Bodo designis* and *Neobodo designis* are earlier and later homotypic synonyms for the same kinetoplastid flagellate. A species concept is the scope of diversity that a name refers to (Franz and Peet 2009). With new insights, concepts may be split (Sonneborn 1975), new names are created, and the meaning of old ones such as *P. aurelia* change or becomes confused. Homonyms are formed when the same name is applied to more than one species concept. Whether through oversight, different stances about lumping and splitting, or because of attention to different life stages, the same species may be described as new by more than one person, each using a new name. When the common identity is asserted, the dif-

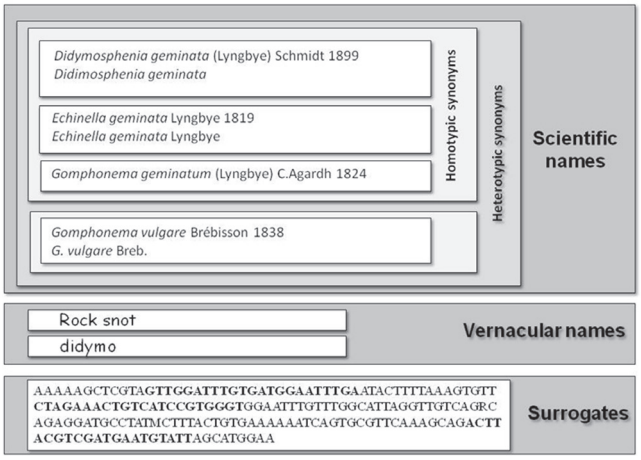

Fig 2. Reconciliation of alternative names for the same taxon (an invasive diatom species). The diagram shows three classes of 'names': scientific names, vernacular names, and surrogates or strings that act in the same way as names (sequence data in this example). *Gomphonema vulgare* and *Echinella geminata* were applied independently to the same species and are heterotypic synonyms. The reconciliation group includes the homotypic synonyms (*Echinella geminata* and *Didymosphenia geminata*), and the lexical variants of all names. Reconciliation groups allow computer-based queries initiated with one name to be answered with information associated with all names.

ferent names are recognized as heterotypic synonyms (Fig. 2). A names-based cyberinfrastructure must overcome these problems so that all information on a species can be joined together, irrespective of the name that was used for it; and that the results should not be contaminated with information about other species with the same name.

Two solutions resolve the 'many names for one taxon' problem – standardization and reconciliation. Standardization requires the use by everyone of the same name-string. It assumes that there can be one consensual name for a taxon. This is not correct because taxonomy is a dynamic discipline (Franz and Thau 2010) involving the judgments of many taxonomists. Standardization cannot be applied to older documents without recourse to reconciliation.

Reconciliation maps alternative name-strings for the same species together (Fig. 2). Reconciliation includes scientific names, vernacular names, and surrogates for names, homotypic and heterotypic synonyms, and lexical variants of all. Reconciliation is an unavoidable component of a names-based infrastructure (Patterson *et al.* 2010). Reconciliation groups can be built in part algorithmically using fuzzy matching (URL 36) to overcome typographic and OCR errors that affect a single or a few characters, and by parsing algorithms to fragment name strings into their component parts and remove particularly unreliable parts such as the name of the author or the date when the name was published. Algorithms that target species epithets, author information, name of the basionym author, and taxonomic context can identify that *Bodo designis* Skuja 1948 and *Neobodo designi*s (Skuja 1948) Vickerman 2004 are homotypic synonyms.

The most significant impediment to providing comprehensive reconciliation services is the absence of synonymy information from on-line sources. Most of that information can only be found on paper.

Valuable services can be built on top of reconciliation. Resolution services use information in reconciliation groups to return the senior synonym from a preferred taxonomic authority (Boyle *et al.* 2013). Resolution can convert names within older documents into current names and normalize the names components of databases. Coupled with names recognition software, resolution services can be turned into taxonomic validation modules for word-processing, spreadsheet or database softwares, or in contemporary publication workflows – checking if names are current and taxonomically endorsed as they are typed in or uploaded to new environments. Such a tool would prevent users from using incorrect names or eliminate the idiosyncratic name strings encountered in Dryad (see above).

HIGHER TAXON NAMES

The names of higher taxa and their placement within hierarchies have under-used potential within information management. They may be used to navigate and browse content or, with access to a compilation of taxonomic information, they can be dereferenced to all component taxa. Derefencing would enable queries that refer to higher taxa to return data about all species in the taxon. These taxonomically aggregative searches have considerable potential in research. They can be used to test hypotheses as to whether a taxon is holophyletic (Ashlock 1971) by establishing if key features are present in all species, or if a taxon is polyphyletc, or paraphyletc. Hierarchies built of holophyletic clades can be used to predict attributes in members of the clade even though no records exist. Classifications based on holophyly are not the only option for biology. We might classify by geography, or by familiar plesiomor-

phic features when we refer to organisms as microbes, parasites, or chromists. Within a digital world, these approaches are not mutually exclusive (Weinberger 2007), but a system that includes an assembly of holophyletic clades will reward us with biologically meaningful data organization and enhanced search performance.

'Names' will remain, at least for a while, our best tokens by which we identify supra-generic clades. As with species, "names can serve an efficient mean of communication only if they are relatively stable over time" (Vences *et al.* 2013). 'Stability' is achieved when the name is unambiguous as to what it means and the meaning does not change. This can be achieved if there is only one name for the clade (i.e. there is no synonymy), and that any name is only ever used to refer to one thing (there is no homonymy). It is unusual for homonymy or synonymy to arise because of ignorance of pre-existing names for supra-generic taxa, but may arise from the practice of claiming authorship of an existing term that Dubois (2006) refers to as dishonest. Problems of synonymy and homonymy are familiar (Patterson 1999), but exacerbated at higher ranks because the inapplicability of the codes of nomenclature adds to instability (Vences *et al.* 2013).

Synonymy is relatively easy to fix through reconciliation. Homonymy, on the other hand, arises when a name has more than one meaning. The meaning of a name is a 'concept' (Franz and Peet 2009). We encounter problems if the concept that is referred to by a name is not clear (e.g. Archaeprotista – Margulis 1996); if a name is used for profoundly different concepts (e.g. Protozoa); or is the meaning of the name meanders (e.g. Archamoebae and Chromista). A simple metric of how well a concept is understood is 'How easy is it to dereference the name to all of its component taxa?.' A concept is stable when that process always leads to a consistent group of species. This is rarely achievable with protists because there is much subjectivism in forming higher taxon names (Lahr *et al.* 2012, Patterson 1999, Wegener-Parfrey *et al.* 2006). Vences *et al.* (2013) discuss how best to minimise subjectivism. They urge for taxa to be holophyletic, for them to have phenotypic diagnosability via synapomorphies, but most importantly urge for caution in creating new taxa "when different monophyly-based classifications are conceivable." That is the essence of the problem within protistology – many taxa are premature – being erected while different monophyly-based classifications are conceivable. The existence of the problem is evident from the short-lived nature of higher names or because the concepts they refer to are not stable (Patterson 1999). Many premature acts have their origins in molecular 'phylogenies.' Despite bootstrapping and other tests, dendrograms from phylogenetic analyses are inherently probabilistic and many conceivable arrangements are possible from single studies. Molecular phylogenies and taxonomies derived from them conflict with the Vences *et al.* principle, and undermine the value of higher taxon names in bioinformatics. A simple improvement would be to articulate the synapomorphies associated with the group (Patterson 1982). If such cannot be identified, the group should remain an un-formalized hypothesis (as with the 'hypochondria' – Patterson and Sogin 1992) and not be included within a classification of holophyletic clades.

The importance of synapomorphies is important for another reason. Stability of names for higher taxa will result if we associate a single concept with each name. There are two styles of defining concepts: ostensive and intensional (Dubois 2012, Franz and Peet 2009). Ostensive definitions are those that refer to the content of the taxon. They may be circumscriptions that specify the characters contained within an envelope, or be compositional and refer to some or all children within the clade. Circumscriptions may be self contradictory ("included taxa may be parasitic or free-living"), and are often not exclusive to the taxon in question. A name that is defined by pointing to children will be destabilized as taxonomic and phylogenetic research leads to new taxa being added to a clade, removed, or merged. Higher-taxon names would only be stable in an ostensive system if a new name was created every time the composition or circumscription of the taxon was changed, or if we refine the name with a pointer to what it means (e.g. Protozoa *sensu* von Siebold 1845 versus *sensu* Cavalier-Smith 1993).

Intensional definitions refers to the properties that are required for something to be included by the definition (Franz and Peet 2009). If our goal is for holophyletic taxa, then such definitions will likely be phylogenetic. That may give the Phylocode (URL 37), with its emphasis on a nomenclature for clades, new relevance. The preference for definitions based on properties and not topology (inherently ostensive) favours the use of synapomorphies. A synapomorphic-oriented approach has proven to be stable with examples of stramenopiles (Patterson 1989), excavates (Simpson and Patterson 1999), and alveolates (a term first used in-house as the significance of the insights published by Gajadhar *et al.* 1991 were becoming clear).

Intensional definitions have the shortcoming that they are not inherently deferenceable. They alone cannot achieve aggregative searches. For a stable and useful system, we need a system in which taxa with intensional definitions are associated with their taxonomic content.

A concept that is ostensive but that points to precisely the same composition as an intensional definition, is a different concept. When these are confounded, meaningless statements such as "'stramenopiles' is an unnecessary recent synonym for the longer established classical 'heterokonts'" (Cavalier-Smith and Scoble 2013) emerge. The quoted statement is true only if the term 'stramenopile' means the same thing as 'heterokonts.' Yet, the definition of stramenopiles ('taxa with evenly-spaced tripartite tubular hairs, or organisms derived from such taxa' (Patterson 1989) is clearly intensional. The definition of 'heterokont' is not apomorphy-based because the inferred condition of unequal flagella can be found in taxa not included in the heterokonts (e.g. *Notosolenus*). Luther (1899) introduced the 'Klasse Heterokontae' for some xanthophytes and raphidophytes only. The defining criteria and the composition have changed with time (Cavalier-Smith and Chao 1996, Andersen 2004), and none match the concept of 'stramenopiles'.

ANNOTATIONS – TRASHING THE 'RUBBISH IN, RUBBISH OUT' ADAGE

Biologists are often reticent to release content because they are uncertain of the quality and are fearful of how the repercussions may affect their reputation. The consequence of this trepidation is that relatively little of known data makes it into the internet-accessible digital world.

A solution to this dilemma is becoming available through open, communal web-based annotation systems for biodiversity data (Morris *et al*. 2009, Tschöpe *et al.* 2012, Wang *et al.* 2009, URL 38). Such systems require elements of data to be identifiable through UUIDs. With appropriate tools, users can add comments to the (UUIDs for the) data, and the comments can then be made visible to authors, federating nodes, or other users. This allows errors to be identified, corrected, new data added, or gaps identified. Annotation systems offer the possibility of continuous, community-sourced, quality enhancement. Such a system would allow misidentifications in GenBank (e.g. Gomez 2013) to be flagged. It would allow the species name of the first named tintinnid (Müller 1776, 1779) that changed from *inquilinus* to '*inguilinus*' by the time it reached World Registry of Marine Species (URL 39), an entry that purportedly was checked by three people, to be corrected. It would allow the three occurrences of the ciliate family Cyrtolophosidae in the current Catalogue of Life (Fig. 3) to be corrected.

MAKING NODES

If protistology is to engage with the big data world, then the key step will be the development of relevant and effective nodes. Active areas of research mostly deal with data born digital, in which data management environments will emerge as part of the natural process, but will need to take responsibility for automated data flow. Two areas that need to include legacy data spring to mind. The first deals with the 'occurrence' records of species in a particular location at a given time because of their relevance to showing and explaining changing distributions of species. The second relates to protist biodiversity, the taxonomic elements of which require access to all nomenclatural and taxonomic literature from Linnaeus' Species Plantarum (Linnaeus 1753) and the first monograph of protozoa (Müller 1773). The following uses the example of such a virtual 'protistiary' to illustrate what might be in a node and what it might take to create one.

There is no comprehensive catalog of protist diversity. This needs to be rectified as awareness of all species affects the quality of taxonomic judgments, the identification of species, and hence the credibility in dependant studies in ecology and phylogeny. An on-line communal protistiary would provide access to complete classifications, alternative points of view, nomenclatural status and pointers to the literature in which taxonomic judgments are made. The site could link to other information such as distributions and phylogenies. By maintaining authoritative coverage of names, the most valuable metadata for indexing and organizing data about species, taxonomists would regain a new relevance by contributing to biodiversity data management (Patterson 2009).

A comprehensive dynamic on-line classification of protists would re-invent the paper-based classifications (Honigberg *et al.* 1964; Levine *et al.* 1980; Margulis *et al.* 1990; Adl *et al.* 2005, 2012) that are variously

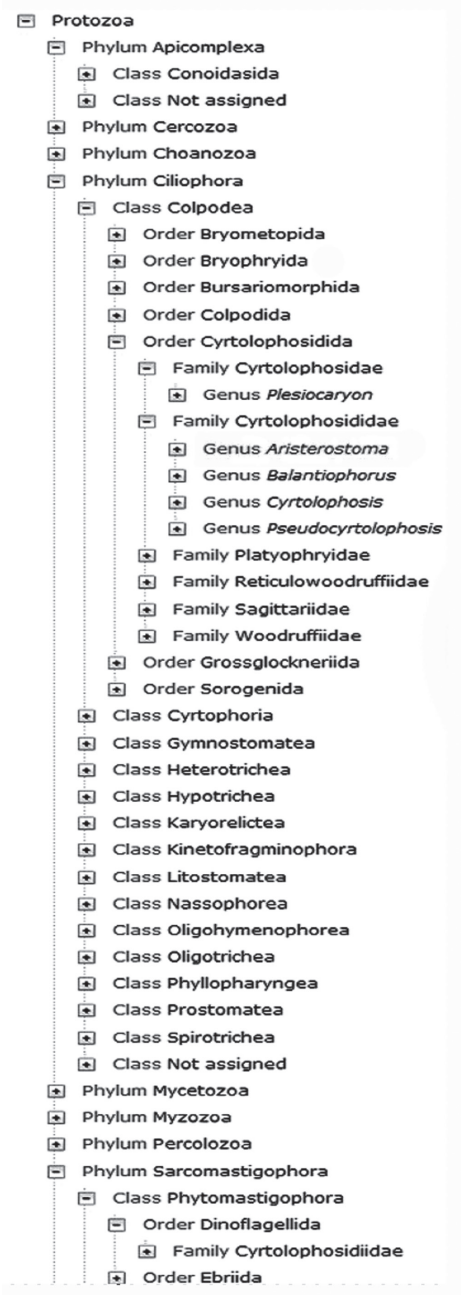

Fig. 3. Selected component of Catalogue of Life (URL 17), showing that the Family Cyrtolophosidae is classified in three locations (with variant spellings).

aggravating by their stasis, incompleteness, principles that allow obsolete and unready components to be included, duplications, and illogicality. The digital world allows a much broader community to engage and to make quantum steps towards higher standards than those set by individuals and committees. Such a system should have the requirement that any taxon/clade name can be expanded to all of its members. Such dereferenceability inevitably includes all plants, animals, and fungi. This requires cross-links that delegate responsibility for information on non-protistan subtaxa to other knowledgeable environments such as Species Fungorum (URL 40), the Plant List (URL 41), the Interim Register of Marine and Nonmarine Genera (URL 42), Catalogue of Life (URL 17), and so on.

Comprehensive also means 'all points of view.' Most taxonomic projects in the digital world do not accommodate multiple points of view. This can be achieved by managing the taxon names independently of the parent-child relationships among the taxa (Pullan *et al.* 2000). The Encyclopedia of Life (URL 14) and iNaturalist (URL 43) illustrate that cross-walks can be built among different co-existing classifications. At the core of a protistiary would be a nomenclator. Nomenclators, inter alia, check that names comply with the appropriate code(s) of nomenclature and provide nomenclatural information. The role of nomenclators will increasingly be fulfilled by on-line registries such as MycoBank (URL 44) and ZooBank (URL 45) (Redhead and Norvell 2012, Pyle and Michel 2008). A nomenclator for protists could be built off the back of ZooBank, but with some new business rules to accommodate algae and ambiregnal taxa (Patterson and Larsen 1991). As Larsen and Patterson (1990) have demonstrated, more than one code can be applied at the same time.

Ideally, nomenclators should point to images of the literature that includes the nomenclatural acts. We need to have all literature that includes nomenclatural and taxonomic acts on line. There are likely to be scuffles over content in publications for which publishers impose 'copyright' restrictions. However, because of their factual nature and their presentation in unoriginal formats, taxonomic treatments are not creative works and are not subject to Intellectual Property restrictions (Agosti and Egloff 2009, Patterson *et al*. 2014) and nomenclatural and taxonomic acts can be extracted for community use. Various on-line environments, such as RefBank (URL 46) and the Biodiversity Heritage Library (URL 47) provide considerable infrastructure that can be used to manage literature and literature citations.

Nomenclators are richer if names are placed within the context of a taxonomic framework with synonymy information. Synonymy information can be used for reconciliation, resolution and taxonomic validation

services. A protistiary would be expected to include reasons for synonymies, link to the relevant literature, offer multiple perspectives, and include communal editing or annotation systems that ensure the completeness and correctness of the inventory. With services that intercept RSS feeds and similar alerts from publishers (Leary *et al.* 2007), or with access to new additions to names registries, the system would remain current with taxonomic advances.

With its access to taxonomic information, a protistiary would provide 'taxonomically intelligent' services. It can offer taxonomies or phylogenies to browse and organize content held at the protistiary or at other locations. It can call on reconciliation to bring together information for the same species even if it has been labelled with different names, be able to present content under taxonomically endorsed names (i.e. have resolution), be current with taxonomic advances, allow aggregative searches so that a query about tintinnids can be exploded into searches for every species of tintinnid and using all names that have ever been used for each species. Such a system needs to be able to deal with homonyms, and should be able to discriminate among concepts. Given the value of synapomorphies, their inclusion for all taxa would improve diagnoses of taxa, allow taxa to be treated as testable hypotheses, and could be exploited by matrix based, identification tools such as Lucid, Delta, X:ID, SLIKS, IdentifyNature, etc. (Dallwitz *et al.* 2007).

ACTION ITEMS FOR CHANGE

Getting into the Big Data world requires investment in technical infrastructure, social change, and in the management of content (Thessen and Patterson 2011). The following checklist offers a guide to key steps in assembling a node:

- A champion identifies an area that will yield rewards (of efficiency or quality) if it is made part of a well-designed piece of infrastructure.
- A proof of concept version of the concept is developed, to help collaborators and users have a sense of what the node might develop into.
- The champion seeks support from colleagues, enemies and users on his or her vision, transforming his or her vision into a community endeavour. The community seeks feedback from primary stakeholders, and petitions the most relevant societies for endorsement.
- A working group emerges to establish the requirements of the infrastructural node, distinguishing essential functions from desirable functions.
- The most similar existing structures are identified, best practices are established, and the costs of achieving the requirements are estimated. Existing wheels in the form of openly available software are identified so that their reinvention can be avoided.
- With some funding, the working group moves towards implementation and the team is expanded to include at least one informatician.
- A prototype version is established at low cost, ideally by modifying existing wheels to meet the needs of the project and to demonstrate the feasibility of the concept.
- A sustainability plan is developed.
- Tumultuous applause from the user community is used to obtain funds that transform the prototype services into production and eventually flawless services.

CONCLUDING REMARKS

Biodiversity informatics is a discipline that is just beginning to form itself. It has enormous potential to become the macroscope that uses information in different domains of knowledge to promote an understanding of the 'infinitely complex' world of Biology (de Rosnay 1975). Disciplines such as protistology have to take responsibility for building their components of this grand machine, and only if we do so, do we get to play on the big stage.

Acknowledgements. I acknowledge support of US National Science Foundation grant DBI-1062387 (The Global Names Architecture, an infrastructure for unifying taxonomic databases and services for managers of biological information), and input from Anne Thessen, Edward Vanden Berghe, John Dolan, and two anonymous reviewers.

REFERENCES

Adl S. M., Simpson A. G., Farmer M. A., Andersen R. A., Anderson O. R., Barta J. R., Bowser S. S., Brugerolle G., Fensome R. A., Fredericq S., James T. Y., Karpov S., Kugrens P., Krug J., Lane C. E., Lewis L. A., Lodge J., Lynn D. H., Mann D. G., McCourt R. M., Mendoza L., Moestrup Ø., Mozley-Standridge S. E., Nerad T. A., Shearer C. A., Smirnov A. V., Spiegel F. W., Taylor M. F. (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* **52:** 399–451

Adl S. M., Simpson A. G. B., Lane C. E., Lukes J., Bass D., Boqwser S. S., Brown M. W., Burki F., Dunthorne M., Hamply V., Heiss A., Hoppenrath M., Lara E., le Gall L., Lynn D. H., McManus H., Mitchell E. A. D., Mosley-Stanridge S. E., Parfrey L. W., Pawlowski J., Rueckert S., Shadwick L., Schoch C. L., Smirnow A., Spiegel F. W. (2012) The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59:** 429–493. doi: 10.1111/j.1550-7408.2012.00644.x

Agosti D., Egloff W. (2009) Taxonomic information exchange and copyright: The Plazi approach. *BMC Research Notes* **2:** 53. doi:10.1186/1756-0500-2-53

Andersen R. A. (2004) Biology and systematics of heterokont and haptophyte algae. *Am. J. Bot*. **91:** 1508–1522

Ashlock P. D. (1971) Monophyly and associated terms. *Syst. Zool.* **20:** 63–69

Baker M. (2010) Next-generation sequencing: Adjusting to data overload. *Nature Method*s **7:** 495–499

Boyle B., Hopkins N., Lu Z., Antonio J., Mozzherin D., Rees T., Matasci N., Narro M. L., Piel W. H., McKay S. J., Lowry S., Freeland C., Peet R. K., Enquist B. J. (2013) The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics* **14:** 16. doi:10.1186/1471-2105-14-16

Caron D. A., Countway P. D. (2009). Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquat. Microb. Ecol.* **57:** 227–238

Caron D. A., Countway P. D., Savai P., Gast R. J., Schnetzer A., Moorthi S. D., Dennett M. R., Moran D. M., Jones A. C. (2009) Defining DNA-based operational taxonomic units for microbial eukaryote ecology. *App. Environ. Microb.* **75:** 5797–5808

Caron D. A., Hutchins D. A. (2013) The effects of changing climate on microzooplankton grazing and commiunity structure: drivers, predictions and knowledge gaps. *J. Plankton Res*. **35:** 235–252

Cavalier-Smith T., Chao E. (1996) 18S rRNA sequence of *Heterosigma carterae* (Raphidophyceae), and the phylogeny of heterokont algae (Ochrophyta). *Phycologia* **35:** 500–510

Cavalier-Smith T., Scoble J. M. (2013) The phylogeny of Heterokonta: *Incisomonas marina*, a uniciliate gliding opalozoan related to *Solenicola* (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. *Eur. J. Protistol*. **49:** 328–353

Chapman A. D. (2005) Principles of data quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-03-8

Chapman A.D. (2009) Numbers of Living Species in Australia and the World, 2nd Edition. Australian Biological Resources Study, Australia

Charvet S., Vincent W. F., Comeau A., Lovejoy C. (2012) Pyrosequencing analysis of the protist communities in a High Arctic meromictic lake: DNA preservation and change. *Front. Microbiol*. **3:** Article 422, 14 pp. doi: 10.3389/fmicb.2012.00422

Dallwitz M. J., Paine T. A., Zurcher E. J. (2007) Interactive identification using the internet. ftp://delta-intkey.com/www/netid.pdf

de Rosnay J. (1975) Le Macroscope. Vers une Vision Globale. Editions de Seuil, Paris

Dubois A. (2006) Proposed rules for the incorporation of nomina of higher-ranked zoological taxa in the International Code of Zoological Nomenclature. 2. The proposed rules and their rationale. *Zoosystema* **26:** 165–258

Dubois A. (2012) The distinction between introduction of a new nomen and subsequent use of a previously introduced nomen in zoological nomenclature. *Bionomina* **5:** 57–80

Franz N. M., Peet, R. K. (2009) Towards a language for mapping relationships among taxonomic concepts. *Syst. Biodiv*. **7:** 5–20

Franz N. M., Thau D. (2010) Biological taxonomy and ontology development: scope and limitations. *Biodiversity Informatics* **7:** 45–66

Gajadhar A. A., Marquardt W. C., Hall R., Gunderson J., Ariztia-Carmona E. V., Sogin M. L., (1991) Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. *Mol. Biochem. Parasitol*. **45:**147–54

GBIF (2011) A beginner's guide to persistent identifiers, version 1.0. http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf

Gómez F. (2014) Problematic biases in the availability of molecular markers in protists: The example of the dinoflagellates. *Acta Protozool.* **53:** 63–75

Gore A. (2013) The Future. Random House, New York

Guillou L., Bachar D., Audic S., Bass D., Berney C., Bittner L., Boutte C., Burgaud G., de Vargas C., Decelle J., Del Campo J., Dolan J. R., Dunthorn M., Edvardsen B., Holzmann M., Kooistra W. H., Lara E., Le Bescot N., Logares R., Mahé F., Massana R., Montresor M., Morard R., Not F., Pawlowski J., Probert I., Sauvadet A. L., Siano R., Stoeck T., Vaulot D., Zimmermann P., Christen R. (2013) The protist ribosomal reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 2013 Jan; 41 (Database issue): D597-604. doi: 10.1093/nar/gks1160

Haeckel E. (1887) Report on the Radiolaria collected by the H.M.S. Challenger during the Years 1873–1876. Report on the Scientific Results of the Voyage of the H.M.S. Challenger, Zoology, Volume XVIII, Her Majesty's Stationery Office, London

Hey T., Tansley S., Tolle K. (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington

Honigberg B. M., Balamuth W., Bovee E. C., Corliss J. O., Gojdics M., Hall R. P., Kudo R. R., Levine N. D., Loeblich A. R., Weiser J., Wenrich D. H. (1964) A revised classification of the phylum Protozoa. *J. Protozool*. **11:** 7–20

Kelling S., Hochachka W. M., Fink D., Riedewald M., Caruana R., Ballard G., Hooker G. (2009) Data-intensive science: a new paradigm for biodiversity studies. *BioScience* **59:** 613–619. doi: 10.1525/bio.2009.59.7.12

Lahr D. J. G., Lara E., Mitchell E. A. D. (2012) Time to regulate microbial eukaryote nomenclature. *Biol. J. Linnean. Soc.* **107:** 469–476

Larsen J., Patterson D. J. (1990) Some flagellates (Protista) from tropical marine sediments. *J. Nat. Hist.* **24:** 801–937

Leary P. R., Remsen D. P., Norton C. N., Patterson D. J., Sarkar I. N. (2007) uBioRSS: Tracking taxonomic literature using RSS. *Bioinformatics* **23:** 1434–1436

Levine N. D., Corliss J. O., Cox F. E. G., Deroux G., Grain J., Honigberg B. M., Leedale G. F., Loeblich A. R., Lom J., Lynn D. H., Merinfeld G., Page F. C., Poljansky G., Sprague V., Vavra J., Wallace F. G. (1980) A newly revised classification of the Protozoa. *J. Protozool*. **27:** 37–58

Linnaeus C. (1753) Species Plantarum. Salvius, Stockholm

Luther A. (1899) Ueber *Chlorosaccus* eien neue Gattung der Süsswasseralgen nebst einigen Bemerkungen zur Systematik verwandter Algen. *Beih. Kongl. Svenska Vetensk. Akad. Handl.* **24 (III 13):** 1–22

Margulis L. (1996) Archael-eubacterial mergers in the origin of Eukarya: Phylogenetic classification of life. *Proc. Natl. Acad. Sci* **93:** 1071–1076

Margulis L., Corliss J. O., Melkonian M., Chapman D. J. (1990) Handbook of Protoctista. Jones and Bartlett Publishers, Boston

Morris P. J., Kelly M., Lowery D. B., Macklin J. A., Morris R., Tremonte D., Wang Z. (2009) Filtered Push: Annotating distributed data for quality control and fitness for use analysis. Eos Transactions of the American Geophysical Union (AGU) 90(52) Fall Meeting Supplement, Abstract available at http://adsabs.harvard.edu/abs/2009AGUFMIN34B..08M

Müller O. F. (1773) Vermium terrestrium et fluviatilium, seu, Animalium infusoriorum, helminthicorum, et testaceorum, non marinorum succincta historia. Havniae & Lipsiae

Müller O. F. (1776) Zoologiae Danicae prodromus, seu animalium Daniæ et Norvegiae indigenarum characteres, nomina, et synonyma imprimis popularium. Havniae. (Hallager)

Müller O. F. (1779) Zoologia danica sev animalivm Daniae et Norvegiae rariorum ac minus notorum descriptiones et historia. Volumen primum. Explicationi iconum fasciculi primi eiusdem operis inserviens. Havniae & Lipsiae

National Science Foundation (2006) NSF's Cyberinfrastructure vision for 21st Century discovery, v. 5.0. NSF Cyberinfrastructure Council http://www.nsf.gov/od/oci/ci_v5.pdf

National Science Foundation (2011) A Report of the National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges. National Science Foundation, Washington, DC, http://www.nsf.gov/od/oci/taskforces/TaskForceReport_GrandChallenges.pdf

National Research Council of the National Academies (2009) A New Biology for the 21st Century. National Academies Press, Washington, DC, http://www.ncbi.nlm.nih.gov/books/NBK32509/pdf/TOC.pdf

Page R. D. M. (2011) Dark taxa: GenBank in a post-taxonomic world. Available at http://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html

Patterson C. (1982) Morphological characters and homology. In: Problems in Phylogenetic Reconstruction, Systematics Association Special Volume 21, (Eds. K. A. Joysey, A. E. Friday). London: Academic Press, 21–74

Patterson D. J. (1989) Stramenopiles: chromophytes from a protistological perspective. In: The chromophyte algae: Problems and perspectives, (Eds. J. C. Green, B. S. C. Leadbeater, W. L. Diver). Clarendon Press, Oxford, 357–379

Patterson D. J. (1999) The diversity of eukaryotes. *Amer. Natur.* **154:** S96–124

Patterson D.J. (2009) Future taxonomy. In: Systema Naturae 250 – the Linnaean Ark, (Ed. A. Polaszek), CRC Press, London, 115–124

Patterson D. J., Egloff W., Agosti D., Eades D., Franz N., Hagdorn G., Rees J., Remsen D. (2014) Scientific names of organisms: attribution, right, and licensing. *BMC Research Notes* (in press).

Patterson D. J., Larsen J. (1991) Nomenclatural problems with protists. *Regnum Vegetabile* **123:** 197–208

Patterson D. J., Sogin M. L. (1992) Eukaryote origins and protistan diversity. In: The origin and evolution of the cell, (Eds. H. Hartman, K. Matsuno), World Sci., Singapore, 13–47

Patterson D. J., Remsen D., Norton C., Marino W. (2006) Taxonomic Indexing – extending the role of taxonomy. *Systematic Biology* **55:** 367–373

Patterson D. J., Faulwetter S., Shipunov A. (2008) Principles for a names-based cyberinfrastructure to serve all of biology. *Zootaxa* **1950:** 153–163

Patterson D. J., Cooper J., Kirk P. M., Pyle R. L., Remsen D. P. (2010) Names are key to the big new biology. *TREE* **25**: 686–691. doi:10.1016/j.tree.2010.09.004

Pawlowski J., Christen R., Lecroq B., Bachar D., Shahbazkia H. R., Amaral-Zettler L., Guillou L. (2011) Eukaryotic richness in the abyss: Insights from pyrotag sequencing. *PLoS One* **6:** e18169

Pierce R. W., Turner J. T. (1993) Global biogeography of global tintinnids. *Mar. Ecol. Prog. Ser.* **94:** 11–26

Piwowar H. A., Vision T. J., Whitlock M. C. (2011) Data archiving is a good investment. *Nature* **473:** 285. doi:10.1038/473285a

Pullan M. R., Watson M. F., Kennedy J. B., Raguenaud C., Hyam R. (2000) The Prometheus taxonomic model: A practical approach to representing multiple classifications. *Taxon*. **49:** 55–75

Pyle R., Michel E. (2008) Zoobank: Developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa* **1950:** 39–50

Raup D. (1991) Extinction: Bad Genes or Bad Luck?, Norton, New York

Redhead S. A., Norvell L. L. (2012) MycoBank, Index Fungorum, and Fungal Names recommended as official nomenclatural repositories for 2013. http://www.imafungus.org/Issue/32/03.pdf

Sarmento H., Montoya J. M., Vázquez-Domínguez Váqué D., Gasol J. M. (2010) Warming effects on marine microbial food web processes: How far can we go when it comes to predictions? *Phil. Trans. Roy. Soc. B* **365:** 2137–2149

Simpson A. G. B., Patterson D. J. (1999) The ultrastructure of *Carpediemonas membranifera* (Eukaryota) with reference to the "excavate hypothesis". *Eur. J. Protistol*. **35:** 353–370

Sonneborn T. M. (1975) The *Paramecium aurelia* complex of fourteen sibling species. *Trans. Am. Microsc. Soc*. **94:** 155–178

Thessen A., Cui H., Mozzherin D. (2012) Applications of Natural Language Processing in biodiversity science. *Advances in Bioinformatics.* doi:10.1155/2012/391574. http://www.hindawi.com/journals/abi/2012/391574/

Thessen A. E., Patterson D. J. (2011) Data issues in the life sciences. *ZooKeys* **150:** 15–51. doi: 10.3897/zookeys.150.1766

Thessen A. E., Patterson D. J., Murray S.A. (2012) The taxonomic significance of species that have only been observed once: The genus *Gymnodinium* (Dinoflagellata) as an example. *PLoS ONE* **7:** e44015. doi:10.1371/journal.pone.0044015

Tschöpe O., Suhrbier I., Güntsch A., Berendsohn W. G. (2012) AnnoSys: A generic annotation system for biodiversity data. GBIF European Regional Nodes Meeting 2012 27.-29.3., Berlin

Vences M., Guayasamin J. M., Miralles A., de la Riva I. (2013) To name or not to name: Criteria to promote economy of change in Linnaean classification schemes. *ZooTaxa* **36:** 201–244

Wang Z., Dong H., Kelly M., Macklin J. A., Morris P. J., Morris R. A. (2009) Filtered-Push: a map-reduce platform for collaborative taxonomic data management. *2009 WRI World Congress on Computer Science and Information Engineering* **3:** 731–735. Available at http://bdei2.cs.umb.edu/wiki/images/a/a8/PID797506.pdf

Wegener-Parfrey L., Barbero E., Lasser B., Dunthorn M., Bhattacharya D., Patterson D. J., Katz L. A. (2006) Evaluating support for the current classification of eukaryotic diversity. *PLOS Genetics* **2:** 2062–2073

Weinberger D. (2007) Everything is miscellaneous. Henry Holt and Company, New York

Wetterstrand K. A. (2013) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts. Accessed 17th March 2013

Wood J., Andersson T., Bachem A., Best C., Genova F., Lopez D. R., Los W., Marinucci M., Romary L., van de Sompel H., Vigen J., Wittenburg P. (2010) Riding the wave. How Europe can gain from the rising tide of scientific data. Final report to European Commission by the High Level Expert Group on Scientific Data. European Union

Zhang W., Feng M., Yu Y., Zhang C., Sun J., Xiao T. (2011) Species checklist of contemporary tintinnids (Ciliophora, Spirotrichea, Choreotrichia, Tintinnida) in the world. *Biodiversity Science* **6:** 655–660. doi: 10.3724/SP.J.1003.2011.06136

URL LINKS

1. US Fair Access to Science and Technology Research Act of 2013: http://beta.congress.gov/bill/113th-congress/house-bill/708?q =hr708
2. GBIF: http://www.gbif.org/
3. OBIS: http://www.iobis.org/
4. Genbank: http://www.ncbi.nlm.nih.gov/genbank/
5. Pangaea: http://www.pangaea.de/
6. Biodiversity Heritage Library: http://www.biodiversitylibrary.org/
7. WoRMS: http://www.marinespecies.org/aphia.php?p=taxdetails&id=183551
8. Tree of Life: http://tolweb.org
9. micro*scope: microscope.mbl.edu
10. The Plankton Ciliate project: http://www.liv.ac.uk/ciliate/intro.htm
11. Villefranche sur Mer web site: http://www.obs-vlfr.fr/LOV/aquaparadox/html/ClassicMonographs.php
12. The Protist Information Server: http://protist.i.hosei.ac.jp
13. Checklist of Phytoplankton in the Skagerrak-Kattegat:http://www.smhi.se/oceanografi/oce_info_data/plankton_checklist/ssshome.htm
14. Encyclopedia of Life: http://eol.org
15. DiscoverLife: www.discoverlife.org/
16. Atlas of Living Australia: www.ala.org.au/
17. Catalogue of Life: http://www.catalogueoflife.org/col/search/all
18. Marine Species Identification portal: http://species-identification.org/
19. ScratchPads: http://scratchpads.eu/
20. Nucleotide Sequence Database Collaboration: http://www.insdc.org/
21. PlasmoDB: http://plasmodb.org/plasmo
22. ToxoDB: http://toxodb.org
23. World Foraminifera Database: http://www.marinespecies.org/foraminifera/
24. AlgaeBase: algaebase.org
25. Biodiversity Information Standards (TDWG): http://www.tdwg.org/
26. DiGIR: http://digir.sourceforge.net/
27. TAPIR: http://www.tdwg.org/standards/449/
28. GBIF IPT: http://ipt.gbif.org/
29. Linked Open Data Cloud: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
30. Open Archives Initiative Protocol for Metadata Harvesting (OAI - PMH): http://www.oaforum.org/
31. Dryad data repository: http://datadryad.org/
32. Marine Metadata initiative: https://marinemetadata.org/
33. TDWG Biodiversity ontologies: https://code.google.com/p/tdwg-rdf/wiki/BiodiversityOntologies
34. IRMNG homonyms: http://www.cmar.csiro.au/ datacentre/irmng/ homonyms.htm
35. Global Names Index: gni.globalnames.org
36. TaxaMatch fuzzy matching algorithm: http://www.cmar.csiro.au/datacentre/taxamatch.htm
37. Phylocode: http://www.ohio.edu/phylocode/
38. Annosys: http://wiki.bgbm.org/annosys/index.php/Main_Page
39. World Registry of Marine Species: http://www.marinespecies.org
40. Species Fungorum: http://www.speciesfungorum.org/
41. The Plant List: http://www.theplantlist.org/
42. The Interim Register of Marine and Nonmarine Genera: http://www.cmar.csiro.au/datacentre/irmng/
43. iNaturalist: http://www.inaturalist.org/
44. MycoBank: http://www.mycobank.org/
45. Zoobank: http://www.zoobank.org
46. RefBank: http://vbrant.eu/content/refbank
47. Biodiversity Heritage Library: http://www.biodiversitylibrary.org/

Received on 12th June, 2013; revised on 19th August, 2013; accepted on 20th August, 2013

Box B. Some terms used by biodiversity informaticians

Aggregate: To bring together information from different digital sources. As an example, the web pages of Encyclopedia of Life, such as http://eol.org/pages/488716, aggregates information from multiple sources.

Annotation: A mechanism that allows additions to be made to digital objects; more particularly annotation systems allow recipients of information to add comments to the information, and for those comments to be returned to the source of the content. FilteredPush (http://wiki.filteredpush.org/) is an example of such a system.

Atoms (data atoms): The smallest effective unit of data, such as a number or other value of a variable.

Atomization: The process of transforming a data file into its component data **atoms**.

Biodiversity informatics: That subdomain of informatics that is relevant to information that relates to biodiversity.

Bioinformatics: The subdomain of informatics that is relevant to molecular biology.

Data model: The conceptual framework to represent information in a particular domain, includes the arrangement of data in tables and the organization of the tables, and mechanisms to acquire and share information.

Data: Factual information that has not been subject to interpretation – also called raw data. May take the form of observations of nature, or from nature distorted as experiments, data produced by running models, or data that are computed from facts. Data are plural.

Dereference: To access the digital information that a pointer, such as a **URL**, **GUID** or **UUID**, points to.

Discipline expertise: With skills and knowledge of the discipline to which informatics may be applied. Protistologists are custodians of expertise in the discipline of protistology.

Discovery metadata: A class of metadata indicating that data on a specified subject can be found at a specified location.

Federate: To bring together data from multiple sources, combining the data and allowing it to be accessed as a single source. OBIS collects data on marine organisms, offering a single point of access to data on the distribution of *Cafeteria* from several sources.

GUID (Globally Unique Identifier): A **standard** sequence of characters **(a string)** that uniquely identifies a digital item that can be accessed through the web, and can point to – direct users to – that item. An example is a Life Science Identifier (LSID) such as http://lsid.tdwg.org/summary/urn:lsid:ubio.org:namebank:2677766. LSIDs are made unique by being placed in the context of a 'web site,' even if the number 2677766 is used by multiple web sites. LSIDs need the context of the web site to find out what is referred to (cf **UUID**).

Identifier: A **string** or other reference that is used as a label for an object, much as a number plate is used to identify a vehicle. The OBIS taxon identifier 414377 refers to the genus *Cafeteria*.

Informatics: The academic discipline of managing digital information, combines expertise in the subject about which information relates and computer sciences.

Information: Data placed in context; whereas 48.3°C may be data, 48.3°C air temperature in the shade in Phoenix Arizona on June 29th 2013 is information.

Knowledge: Knowledge is a consensus that emerges as information is ordered and rationalized – such as, the temperature in Phoenix on June 29th 2013 is the highest on record.

Legacy data: Data not born digital, such as observations made in the 19th century and content of the older literature.

Metadata: terms that are used to describe data, they may define the nature of the data, the organization of the data file, who produced the data, or describe the data (the °C associated with the value 48.3). Increasingly, there are agreed **standards** for metadata, and this improves exchange of information by computers. Metadata are often organized in a logical and formal structure called an **ontology**.

Nomenclator: Individual or organization concerned with the nomenclature of organisms, identifying compliance with appropriate codes of nomenclature, listing and organizing information about code-compliant nomenclatural acts.

Normalize: To eliminate inconsistencies of units or other structuring elements associated with data or digital files, such as transforming all units of length to the metric system, or ensuring that all references to temperature use Centigrade. Normalization is a necessary step to make information useful. It is usually associated with **standardization** (i.e. normalize in an agreed way).

Ontology: A formal structure that is used to declare the relationship between concepts, which in the world of informatics are represented by **metadata** terms or terms from **vocabularies**. Ontologies are used to organize information and to 'represent knowledge' in a way that is understandable to computers.

Open: Refers to software or data, and is a philosophy that contrasts with a commercial approach and in which the software, what it does, or the information it works with is available without payment. Open source content may require that the source of information is identified. Open content is usually accessible without charge (i.e. is free).

Parent child: Refers to an arrangement of information in databases that can be used to represent a hierarchy, in which one object (the child) is identified as being a member of a group in which all members share a relationship with another object (the parent). This structure is well suited to representing biological classifications.

Pointer: A string or identifier that is used to identify a digital object such as a web page. Pointers include **URL**s, **GUID**s and **UUID**s, and are usually intended to be understood and used by computers.

Reconciliation: Linking all known name(-strings) for a taxon together so that a query initiated by one name can be expanded to actions involving all names.

Resolution: The conversion of one name for a taxon to another that is deemed to be correct by a taxonomic authority.

Services: Actions, such as the exchange of data, that are mediated by computers.

Standardize: Make compliant with industry **standards.**

Standards: Agreed formats as to the way **data** and **metadata** may be organized, can include but not be limited to the units used for data, the use of terms for metadata and their **ontology**, the formats of files in which data are compiled or exchanged.

String: A sequence of alphanumeric characters (letters, numbers and other symbols), spaces, and punctuation.

Taxonomic intelligence: Inclusion of specialist taxonomic knowledge (such as synonymies) in biodiversity informatics tools and services.

Taxonomic validation services: Services that check names against reference systems to ensure that they are spelled correctly, have correct authority information, and are endorsed by a taxonomic authority.

Truth: Consensus as to the interpretation of information (knowledge).

URL (Uniform Resource Locator): a universally recognized standard that uses a string composed in a particular way to point to a location that is accessible through the internet (http://www.eko.uj.edu.pl/ap/).

UUID (Universally Unique Identifier): A unique string that complies with agreed standards to identify, typically, a digital item such as datum point, a web page, a body of knowledge. It can be presented in various forms, but ideally includes a 32 digit number. And example is urn:lsid:zoobank.org:pub:E4CF5F07-EC08-4D5D-943C-2E92EBC7D67E. To the right of the last colon is the unique 32 digit 'number,' prior to that is information as to who minted (created) the number, and what it relates to. A UUID does not need the context of a web site.

Vocabulary: A suite of agreed terms used as metadata or to represent concepts in an ontology or actions as in annotation.