The big questions for biodiversity informatics

by A. Townsend Peterson, Sandra Knapp,
Robert Guralnick, Jorge Soberón and Mark T. Holder

2010

Peterson, T., Knapp, S., Guralnick, R., Soberón, J., Holder, M. 2010. The Big Questions For Biodiversity Informatics. Systematics and Biodiversity 8(2): 159-168.

Published version: http://dx.doi.org/10.1080/14772001003739369

Systematics and Biodiversity (2010), **8(2)**: 159–168

Taylor & Francis
Taylor & Francis Group

Perspective
The big questions for biodiversity informatics

A. TOWNSEND PETERSON[1], SANDRA KNAPP[2], ROBERT GURALNICK[3], JORGE SOBERÓN[1] & MARK T. HOLDER[4]

[1]Biodiversity Institute and Department of Ecology and Evolutionary Biology, The University of Kansas, Lawrence, Kansas 66045, USA
[2]Department of Botany, The Natural History Museum, Cromwell Road, London SW7 5BD, UK
[3]Natural History Museum and Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309, USA
[4]Department of Ecology and Evolutionary Biology, The University of Kansas, Lawrence, Kansas 66045, USA

(Received 2 February 2010; revised 28 February 2010; accepted 1 March 2010)

Science is a sequence of generating new ideas, detailed explorations, incorporation of the results into a toolbox for understanding data, and turning them into useful knowledge. One recent development has been large-scale, computer-aided management of biodiversity information. This emerging field of biodiversity informatics has been growing quickly, but without overarching scientific questions to guide its development; the result has been developments that have no connection to genuine insight and forward progress. We outline what biodiversity informatics *should be*, a link between diverse dimensions of organismal biology – genomics, phylogenetics, taxonomy, distributional biology, ecology, interactions, and conservation status – and describe the science progress that would result. These steps will enable a transition from 'gee-whiz' to fundamental science infrastructure.

Key words: analysis, biodiversity data capture, data integration, ecology, evolutionary biology, informatics, interpretation, phylogeny

The questions

Biodiversity informatics is a young field, with the earliest citation of the term only 12 years ago (Schalk, 1998). As a consequence, the field is still at an early stage in its development, and is evolving and extending, and developing its own literature and its own frameworks. However, in our view, the field appears to be growing in a void of overarching, motivating scientific questions, effectively making it a set of technologies in search of questions to address. That is, unlike other relatively new fields (e.g. phyloinformatics), biodiversity informatics currently exists without major, guiding goals that represent intellectual frontiers and challenges. This gap, we fear, leaves the field without a framework for effective thinking.

We do not purport to identify *the* questions for biodiversity informatics as a field in this commentary. Rather, we hope to propose ideas that can stimulate discussion of what these questions can be. Perhaps the broadest suite of goals for the field is not particularly difficult to envision ... something along the lines of *understanding the behaviour of evolving lineages across time, space, phenotype, genotype, environments, and biotic interactions* ... in essence, the challenge of understanding 'space, time, and form' (Croizat, 1962; Nelson & Platnick, 1981). This statement, however, is so broad as to be almost uninformative, as it covers much of biology!

As a consequence, we offer this commentary on the present and future of the field of biodiversity informatics, as well as on its potential to grow and extend strategically as a discipline of inquiry. We envision biodiversity informatics as a vibrant field of inquiry with the potential to 'nucleate' new ideas and novel insights into some of the most interesting and important current challenges in organismal biology, going farther afield even than the big questions in systematics (Cracraft, 2002). Nonetheless, we see the current lack of guiding conceptual frameworks as retarding the advance of the field, with the potential to sideline it completely in terms of serious scientific relevance, ultimately turning it into a series of technological initiatives whose utility will forever be in the future. We hope that those researchers in related fields, who share a vision of a rich conceptual framework for biodiversity informatics, will step up to the challenge, and recast this young field in a rich framework of ideas and concepts.

Correspondence to: A. Townsend Peterson. E-mail: town@ku.edu

ISSN 1477-2000 print / 1478-0933 online

DOI: 10.1080/14772001003739369

Current state

Activities in biodiversity informatics are currently data-centred, falling into three broad categories: (1) data extraction and capture, (2) data compilation and serving and (3) data display and visualization. Data extraction projects include digitizing specimen records, and sorting through digitized literature for taxon names or geographic locations. Data compilation efforts include tasks such as assembling communities of data owners, organizing their information, and publishing it to biodiversity information networks (e.g. GBIF, REMIB, VertNet, SpeciesLink; Graham *et al.*, 2004; Stein & Wieczorek, 2004; Guralnick & Hill, 2009). In parallel, major effort has been invested in compiling digital taxonomic name resources and the taxonomic literature (Koning *et al.*, 2005; Sautter *et al.*, 2006). Finally, display and visualization efforts include presenting compiled data on maps, or creating information pages with biodiversity data 'mashed up' with other types of data (Butler, 2006; Janies *et al.*, 2007).

The goals of digitizing, aggregating and displaying biodiversity data and information are essential, and creation of larger and larger stores of global biodiversity information remains an important locus of biodiversity informatics as a discipline. However, we believe that collection of data and creation of tools for their own sake, and not in service to a clear *need for* those data or tools among the community of biodiversity researchers, managers and decision makers is perilous – as a consequence, careful assessment of the needs of the community of potential users in biodiversity science should be a critical precursor to any and all tool development in biodiversity informatics. This gap separating data and tools from conceptual frameworks pervades every aspect of the biodiversity informatics endeavour. Provision of biodiversity data and development of tools without clear connection to important research questions proposed by end users is nonetheless rampant in this emergerging field.

A clear example of this problem is the absence of adequate metadata associated with biodiversity data. As will be discussed below, rich record-level metadata are key in enabling detailed science applications of biodiversity data. The biodiversity informatics community is finally beginning to embrace the need for better data description through summary reporting, community annotation, and better capture of metadata, to assist all users in determining data fitness of use. However, this data enrichment process has lagged far behind other efforts. Attempts to filter and use biodiversity data effectively to address real questions have already been hampered as a consequence (e.g. Yesson *et al.*, 2007).

Ultimately, we believe that data need to be collected, integrated and served in the direct service of key questions in evolutionary and environmental biology. Such an inversion of the present biodiversity informatics paradigm (Fig. 1) not only focuses attention on what kind of data to collect,

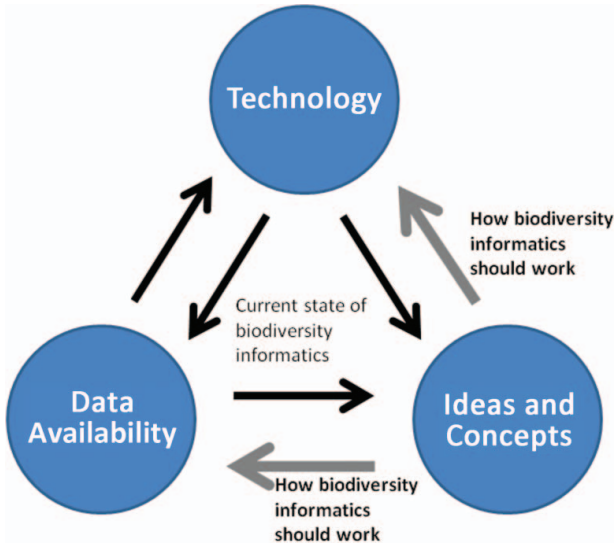

Fig. 1. Illustration of major tendencies in biodiversity informatics: to date, data availability and technology have driven many of the ideas and concepts in the field. The desirable state of the field would see ideas and concepts driving development of new technology and new data resources.

but also on how those data should be processed and described for most effective use by the communities served by biodiversity informatics. Without a strong conceptual framework, we believe that biodiversity informatics may miss its golden opportunity to serve the very communities that need it and motivated its birth. Our impression is that the field presently is developing mostly technology-oriented, 'gee-whiz' gadgets. These flashy – but ultimately empty – products impress the eye, but when the challenge is real production of knowledge, they serve few but the 'inside' practitioners and developers of the field, and do little to advance question-driven science.

A final major problem with the current state of biodiversity informatics is that, with rare exceptions, the field has not effectively linked its fundamental datasets with those from other life and earth science informatics disciplines, whether they be genetic and genomic data repositories, phenotypic information resources, conservation status databases, human information resources, palaeontological data, or rich sets of geospatial environmental data. Present-day informatics systems, for all their ingenuity, are still not particularly capable of tracking disparate data sources and relinking them in meaningful ways; biodiversity data are particularly 'stovepiped' in comparison with others of these fields (Kim, 2002; Koslow & Hirsch, 2004; Hamid *et al.*, 2009). Desperately needed are key informatics products that place biodiversity data in a broader framework linked to other relevant worlds of information.

A panorama of what could be

Given the above summary of how biodiversity informatics currently exists, we now take a look at how it *should*, in

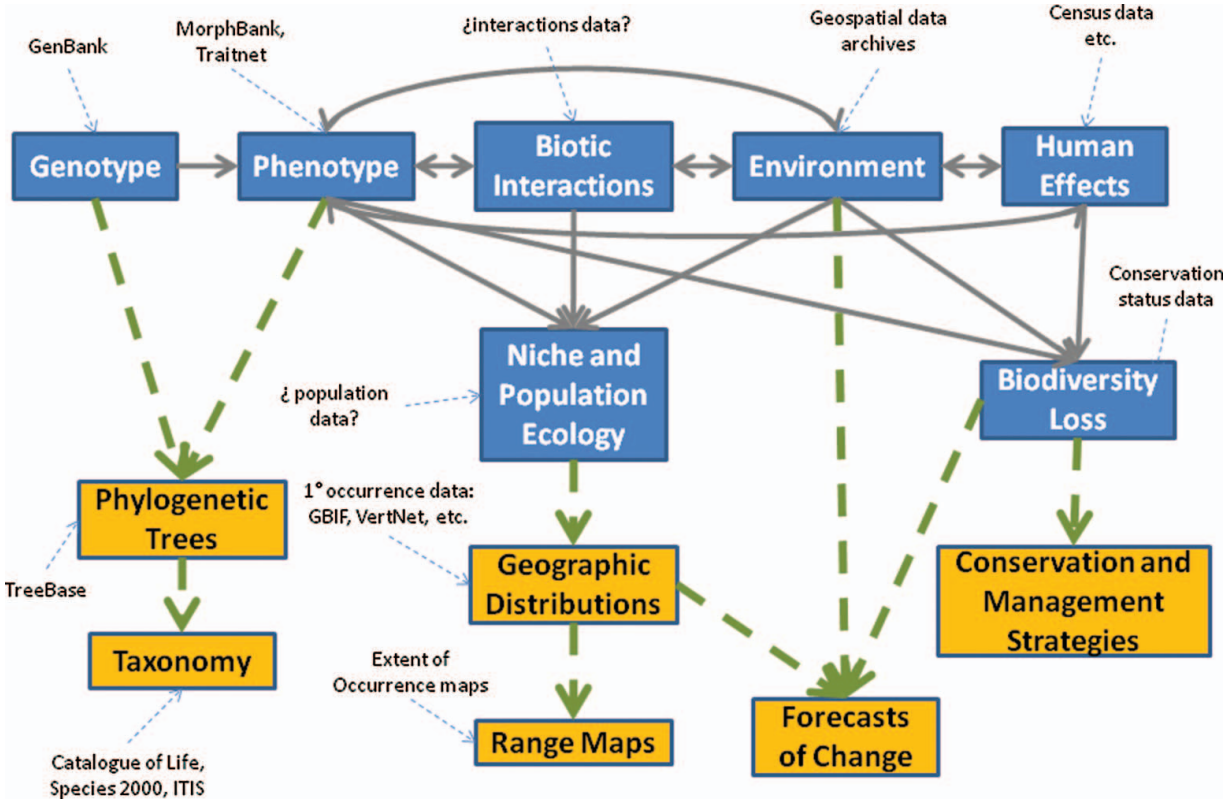

Fig. 2. Summary in broadest terms of the world of biodiversity informatics. Blue boxes are the basic underlying biological processes, ranging from genotype and phenotype up through ecology and biodiversity loss. Orange boxes are the biodiversity information products that are often explored in the field. Labels outside boxes show example information resources or initiatives for most of the elements of the diagram.

our opinion, be organized and structured. The key is this idea of structure and integration – at present, different sectors of biodiversity informatics are largely independent in their development and integration, and cross-linking among sectors has been only cursory and simple. Fig. 2 outlines a broader picture of biodiversity informatics data realms, and shows the potential for rich cross-linkage among presently isolated realms.

The development of geographic information systems (GIS) as a field of inquiry offers an intriguing template for the future evolution of biodiversity informatics. The earliest reference to GIS that we could find in Web of Science was a paper by Grayman *et al*. (1975), who discussed hydrological simulation and analysis. Now, more than 4700 papers later (as indicated in a Web of Science search on 'geographic information system' or 'GIS'), GIS has evolved into a key enabling technology that promotes novel thinking and analyses in fields as diverse as sociology, political science, public health, biogeography and agriculture (Manso & Wachowicz, 2009). What was initially 'just' a software tool became the basis for manifold thinking innovations (including numerous advances in biodiversity informatics) and new paradigms in quantitative geography.

The goal of biodiversity informatics, within the bounds of the overarching question stated at the beginning of this commentary, should be that of enabling new knowledge, new thinking and indeed new questions in biodiversity science. Biodiversity science in general can and should evolve from a purely descriptive cataloguing endeavour into a predictive, scientific exploration of space, time and form. Only in this future and much-different world might biodiversity informatics realize its big-picture potential to enable new science. To illustrate this potential, below, we list five qualitatively new areas of analysis that would be made possible were the level of cross-linking and integration depicted in Fig. 2 to be achieved.

I. Geography and ecology of past life

Documenting the geography and ecology of past life is an essential task in global change biology. Given the current rate and magnitude of environmental change, such documentation takes on special importance in anticipating biotic responses to coming change (Thomas *et al.*, 2004). Our current knowledge of how natural systems are changing remains shallow, despite the pressing need for better understanding of causal factors and processes involved.

How intrinsic factors (e.g. phenology, dispersal capability, habitat specialization and physiological tolerances) and extrinsic factors (e.g. rate and magnitude of abiotic environmental changes, presence and abundance of other species) interact to determine species' responses to environmental change remains little known. A data- and concept-rich infrastructure underpinning global change biology is a fundamental challenge in a century likely to be dominated by environmental change and biodiversity loss.

We believe that a strong conceptual framework for global change biology could be developed via linking of theory and practice in spatial ecology, field ecology and genetics. In each area, theoretical and practical developments such as ecological niche modelling, occupancy modelling and coalescent approaches provide new means to examine changes in species and population distribution and diversity over time. These conceptually diverse tools are now being combined to provide multiple lines of evidence related to global change biology (e.g. Martínez-Meyer *et al.*, 2004, showing integration of palaeontological and recent occurrence data along with palaeo- and recent climate reconstructions). Combination of spatial ecological with phylogenetic–phylogeographic methodologies has proven to be especially fertile ground for new learning (Carstens & Richards, 2007; Knowles *et al.*, 2007; Waltari *et al.*, 2007; Hickerson *et al.*, 2009).

Biodiversity informatics should sit centrally in this only-now-emerging discipline, because the data and tools needed to address global change biology questions span multiple domains and disciplines. Building on the first steps cited above, the potential exists to reconstruct likely distributions of and associations among faunas and floras well back into geological time, yielding a rich picture of how environmental change affects biodiversity and extending the mission of systematics into environmental and ecological realms (Cracraft, 2002). To accomplish development of such a data- and concept-rich framework, we need sources of information regarding past and current geographic distributions of species, phenology, traits and physiology, and detailed pictures of past and present environments. More importantly, we should prioritize development of approaches that facilitate synthesis of these datasets to generate new knowledge regarding past and current changes and their effects on species.

II. Biota-wide picture of diversification and interactions

Community ecology treats the complex realized and potential interactions among elements of biodiversity (Ricklefs & Schluter, 1993). However, its development has suffered from limitations centred around identification of interactants, that is, most of community ecology has focused on the obvious interactions, such as close phylogenetic relatives or known mutualists (e.g. the fungi and algae or cyanobacteria that constitute lichens). However, competition, mutualism, parasitism and other interactions do not necessarily occur only among close relatives, nor have all of these interactions already been noted and recognized by scientists. As such, a broader, more objective approach is both warranted and possible.

Speciation modes saw just such a sea change in approach and insight with a methodology published by Lynch (1989), in which geographic and phylogenetic information was integrated to hypothesize the geography of speciation on an objective basis. Although highly controversial (e.g. Losos & Glor, 2003), this method nonetheless sparked a series of forward steps in objective study of modes of speciation (Fitzpatrick *et al.*, 2009; Kozak *et al.*, 2009). We see a mature version of biodiversity informatics as presenting many opportunities for a much richer and more integrative view of both biotic interactions and biological diversification together, again broadening the mission of systematics considerably (Cracraft, 2002).

Phylogenetic frameworks would be the first ingredient, providing information regarding the evolutionary underpinnings of pairs of taxa. Geographic distributions and niche estimates for species represent additional key elements, showing the relationships of those same taxa in geographic space (across diverse spatial scales) and in environmental dimensions. Finally, new approaches for estimating past distributional patterns (see above) can be marshalled to alleviate problems with labile distributions and the confusion that they insert regarding spatial relations between taxa (Losos & Glor, 2003). The result would be a view of phylogeny and distribution that incorporates information offered by ecological dimensions (Peterson *et al.*, 1999) and likely shifts in distributions over recent planetary history (Wells, 1983). Such a thinking framework could simultaneously offer novel views into interactions among very diverse species *and* the geography and ecology of biological diversification.

III. Future (novel) communities

Palaeontological evidence shows that, in the past, combinations of species existed that have no known extant counterparts (Jackson & Overpeck, 2000). These assemblages are called no-analogue communities (Ackerly, 2003; Williams & Jackson, 2007); their existence suggests that future climate change may produce combinations of species not previously experienced. New communities are appearing as species shift across landscapes tracking changing climates, and reflecting mixtures of native and alien elements; these novel assemblages represent communities for which details of their functioning has not been investigated. The extent to which eradication programmes aimed at non-native elements are successful may be predictable using new information about these no-analogue communities.

To predict no-analogue communities, we must be able to link spatially explicit estimates and forecasts of change phenomena to estimates of ecological niches of current species (both native and alien). Fundamental niches would ideally be estimates directly via experimental or mechanistic approaches (Porter *et al.*, 2002; Kearney & Porter, 2004), but can be estimated under certain scenarios via ecological niche modelling (Soberón & Peterson, 2005). Once niche estimates are in hand, they are integrated with environmental change scenarios to estimate likely future distributions (Jackson & Overpeck, 2000), which are functions of potential niches (the portion of the fundamental niche actually represented under a given current climate, which obviously may be a much smaller subset of the fundamental niche; Soberón & Nakamura, 2009). Finally, potential niches are projected geographically to obtain potentially suitable regions in the planet, and scenarios of dispersal used to estimate likely future distributional areas. A growing number of studies has explored the applicability of these approaches for different sectors of biodiversity (Peterson *et al.*, 2002; Araújo & New, 2007).

A more mature biodiversity informatics, however, would achieve this integration much more broadly with much less effort. At present, species occurrence, habitat data and present-day and future climate data are stored in very distinct stovepipes. What is more, in the case of most climate data, storage formats are quite unfamiliar and inimical to most workers in the biodiversity world. As a consequence, incorporation of new climatic data in biodiversity informatics applications is not at all convenient, and rather the field tends to rely on established sources (Hijmans *et al.*, 2005). Biodiversity informatics tools – properly designed and implemented – could make this process considerably more convenient, not to mention offering it much-improved flexibility and insight.

IV. Integrating phenotype and genotype

Living things are not continuously variable, that is, nature itself is lumpy, with coalescences of characters in what we call taxonomic groups such as species. New data on levels and distributions of diversity, not only in gene sequences but also in regulation patterns of genes, have opened new windows on how nature itself is structured. Traditional taxon circumscription relied on phenotype alone, but has seen considerable refinement of characters and character-states, and presently incorporates data from molecular genetics. Wholesale replacement of phenotypic delimitation with one based solely on DNA sequences has been advocated by some (Vogler & Monaghan, 2006), but natural selection works on organisms' phenotypes, which are products of gene diversity and expression, so the usefulness of the phenotype in understanding biodiversity and describing interactions has refused to go away.

Biodiversity informatics has the potential for integrating phenotypic and genotypic information in novel ways, providing powerful tools for discovery of overlapping and interdigitating patterns in nature. Linking rich data sets on phenotypes and genotypes of individuals, populations and species could be integrated over space and visualized in several dimensions. This approach could provide a unique 'macroscopic' view (Ausubel, 2009) of variation, in which researchers can see – basically for the first time – how genotype and phenotype interact with geography and ecology. This integration will depend on ramping up interaction and collaboration across diverse parts of the scientific community, and appreciation of a wide variety of data types, from ecology and morphology through the regulation of gene expression. The avalanche of data from genomics itself presents an informatics challenge (Roos, 2001; Sonnhammer, 2004), but taking these data beyond a relatively few model organisms and into natural systems will depend upon new biodiversity informatics linkages and workflows. An integrated, multi-dimensional view of variation, with detailed ontologies providing concept-level linkages across domains, will have profound impacts on many other basic science questions.

V. Synthetic conservation planning

Current protected areas for biodiversity conservation have been established – in large part, at least – without explicit consideration of biodiversity patterns (Gotmark & Nilsson, 1992). Although several researchers have developed tools for prioritization of areas under biodiversity criteria (Sarkar *et al.*, 2006), application of these tools to real-world challenges has been rare (see Koleff *et al.*, 2009 for an exception). More recently, however, prioritization thinking has been broadened considerably to include multi-factor, multi-temporal scenarios, which have the potential to cover much more of the true complexity of the challenge (Sarkar *et al.*, 2006).

In a more mature, integrated biodiversity informatics, conservation planning would take advantage of up-to-date, modern taxonomic information (resulting from the integration of genotypic and phenotypic data also facilitated by biodiversity informatics) for careful definition of the units of biodiversity that are the targets of conservation; information on geographic distributions and their likely future configurations (as consequences of land-use change, climate change, species invasions, etc.) to assess the spatial dimensions; and conservation status data and estimates of phylogenetic uniqueness to prioritize particular taxa. All of these elements can be integrated into a more synthetic and robust view of conservation and natural resources management strategy. However, this broad and deep integration can only function to the point that

biodiversity informatics provides sufficient integration and cross-linking.

Key next steps

As discussed above, a fully integrated field of biodiversity informatics would allow us to model jointly the evolution of lineages, ecological niches, geographic distributions and biotic communities by harnessing the tools and data of ecological niche modelling, climate modelling, phylogeographic and phylogenetic methods and other disciplines. Joint estimation of the history of a biota would be much superior to the current approach, which only cobbles together marginal estimates of each aspect (geographic distribution, distribution in environmental space, phylogenetic history, etc.) independently, because a joint estimation approach would allow assessment of uncertainty in estimates at every level of analysis. To enable such integrated inference, however, we must confront some basic concerns about the reliability and combinability of the data on which such inferences would be based.

Data acquisition often constitutes a rate-limiting step in the process of answering scientific questions. Most biologists think about their work as a process of asking questions, designing experiments and then collecting data to arrive at conclusions. Biodiversity informatics, in contrast, presently is facing a different challenge: massive amounts of relevant data are now available via internet portals, but the quality of the data and relationships between distinct data elements are not always apparent.

Hence, a necessary first challenge in maturing biodiversity informatics is *data integration across disparate databases*. For example, we might wish to use data from GenBank or TreeBase as sources of phylogenetic data or MorphBank for phenotypic data for a group of taxa, and use the GBIF or VertNet portal to obtain locality information for those same organisms. A first and critical level of integration depends on a vibrant linkage with alpha taxonomy, that is, species' names and identifications of specimens must be kept current, such that linkages by scientific name are always meaningful. Further and more intimate linkages require unique identifiers at various levels by which to link data records in the two data realms (Page, 2008). Not only do unique identifiers allow fundamental connections between entities referenced in different databases, but associating each datum with a stable, unique identifier makes it possible to study the provenance of the data. Data provenance will continue to be a crucial aspect of the future of biodiversity informatics, because different data sources have varying degrees of reliability, and we frequently use an estimate from one analysis as an input into a subsequent analysis.

Biodiversity informatics is a *sine qua non* for an emerging approach to science that has been called 'Data Intensive Science' (Hey *et al.*, 2009). In more traditional approaches, patterns, processes and regularities are subsumed by theoretical constructs that often take the form of simple equations with a few parameters, like Kepler's or Maxwell's equations. The parameters and structure of these equations encapsulate successfully a very significant part of what it is known about important phenomena. In the data-intensive approach, in contrast, thousands of digital objects comprising terabytes (perhaps petabytes) of data represent the relevant information for the problem, which is impossible to encapsulate in a few simple parameters.

Organizing, visualizing and comprehending the bulk of the data become significant challenges in computer science. An example will illustrate the difference between the two approaches. Calculating the number of species as a function of the area of the region in which they are distributed is a classical problem in ecology. The Arrhenius equation, or SAR, exemplifies the conventional approach: $S(A) = kA^z$, where S is the number of species, A is the area of the region, and k and z are constants. Area and species number are related simply, and k and z are expected to capture the myriad of factors of history, ecology, geography, etc. that translate into numbers of species, and variance around this general relationship is tolerated as noise.

The data-intensive approach, on the other hand, would model the area of distribution of each species individually, by whatever means. Call the estimated area of distribution of species i in region X, $\mathbf{G}_i(\mathrm{X})$. Overlaying all of the distributions of all relevant species predicts (assuming no interactions) the number of species occurring at a given point. In other words, the number of species in the area is the union of all the areas of distribution, or $S(\mathrm{X}) = |\cup \mathbf{G}_i(\mathrm{X})|$, where the bars denote the number of elements of a set. In this approach, instead of two parameters, one gets hundreds, since every species is modelled individually, perhaps using a niche-based approach (Gotelli *et al*., 2009) which is, in itself, a data-intensive endeavour. The equation above therefore actually represents an extremely complicated object, in direct contrast with the quite-simple SARs in the literature.

However, once the entire procedure is defined rigorously, it can be repeated, ideally using scientific workflow software. Most importantly, this approach dispenses with the *ceteris paribus* assumption that the distributional history of large numbers of species will be encapsulated simply by area, and that all other things will be equal. Other things are not equal, and full use of existing data is critical. These force-of-knowledge approaches will thus require quite a different philosophical approach to science.

Data quality and associated signal-to-noise ratios will be of overriding importance in the future of biodiversity informatics, because data documenting many aspects of organisms can be highly heterogeneous in quality. For example, in geographic dimensions, some species descriptions are accompanied by range maps that are drawn freehand on the basis of few records and only rough knowledge of the extent of the species' geographic distribution. In other cases, detailed surveys based on precise geographic references produce information not just on the species' occurrences but also concerning abundance patterns and sites of probable absence. Range maps for species of conservation concern can be extremely precise, in some cases even identifying the locations of all known populations of the species. While all range maps contain some information about species' distributions at some spatial extent and resolution, it is clear that not all range summaries should be accorded the same weight and confidence in analyses of species' distributional patterns. Similar considerations accompany taxonomic dimensions, such that taxonomic information accompanying data records is current, and reflects authoritative naming. If every datum were labelled specifically with a unique identifier and rich metadata documenting quality and precision, the reliablity of downstream estimates could be assessed much more readily.

In the context of data combinability, a detailed understanding of the provenance of data is a crucial prerequisite for *detecting errors and avoiding pseudoreplication*. Issues of pseudoreplication arise in many sectors of biodiversity informatics, particularly as researchers attempt to take advantage of informatics approaches to assemble meta-analyses and other not-originally-planned analyses. For example, it is clearly inappropriate to treat separate inferences of the ecological niche of a species as independent pieces of evidence if they are based on the same set of occurrence data, or to use duplicates of the same botanical collection as independent data points for population assessment.

In some cases, conclusions that would be drawn from combined data if we were to have rich metadata and unique data record identifiers can differ fundamentally from the conclusions we would draw if we do not have access to such data. Consider a species for which a single record places it in an atypical location. If we combine this observation with a population genetic study of the same species that indicates high levels of sequence diversity, the two studies seem to corroborate each other and imply that the species is wide-ranging and under-sampled across its range. If, on the other hand, we could follow the trail of the data back to the original specimens, we might perceive that the specimen in question is an outlier in terms of the ecology and distribution of the species and is genetically very distinct, and the specimen could be considered a misidentification or undescribed taxon. Thus, combining data allows researchers to question the results of the separate analyses based on independent lines of evidence. Developing analytical pipelines for reliable inferences about biodiversity will require error models for the data; to apply these models optimally, we must be able to identify and track the raw data as close to their sources as possible.

Fundamentally, we argue that the record-level metadata associated with biodiversity datastores must be made considerably richer to allow biodiversity informatics to reach its full potential. This message is not new, but is likely to remain a challenge for the field. Completing very rich descriptions of data (and hence rich records with abundant metadata) can slow down a researcher attempting to publish scientific contributions, creating a tension between biodiversity science and the broader needs of the informatics field. More rigorous standards in the biodiversity informatics community for how rich data must be in order to accompany publication could certainly help, and would not impose an unrealistic burden to rapid publication. Promising efforts such as the Dryad repository (http://datadryad.org) attempt to make it easy for researchers to publish rich data relevant to their journal publications.

Finally, a key challenge will be to *deal effectively with scale* throughout biodiversity informatics. For example, a burgeoning field of inquiry is that of macroecology, finding global or regional correlates of species diversity and endemism (Jetz & Rahbek, 2001; Rahbek *et al.*, 2007). These studies, however, are invariably based either on polygon-format summaries of species' ranges (Ridgely *et al.*, 2005) or on coarse grid-based summaries of distributions of species across continents (Rahbek *et al.*, 2007). Although possibly permitting insights into biodiversity patterns and processes at resolutions *coarser than* the already-coarse resolution of the base distributional data sets, these analyses are constrained never to descend to the finer resolutions that – we would argue – are much more relevant to actual organismal biology. This limitation reflects a direct, causal effect of the availability of data on the sorts of questions being asked in one area of biodiversity informatics. At the other end of the scale, the data being assembled using DNA barcodes could, if combined effectively with geography and morphology, allow unique insights into the distribution of genetic diversity on a microscale.

To deal with these scale questions, shifts are necessary in the data available in biodiversity informatics. We suggest a focus on primary data at all times, with secondary data products (e.g. range maps or extent of occurrence maps) being flagged carefully as such. Primary biodiversity data consist of records that place a particular organism (or taxon) at a particular site on Earth at a particular time (Soberón & Peterson, 2005). Extent of occurrence maps and gridded range summaries, in contrast, are secondary information products that are in some way derived from primary data but assume a particular base resolution. Primary records will vary in their base resolutions, depending on the precision with which the geographic referencing was estimated – careful documentation of this resolution in

the form of record-level metadata detailing how the record originated, and how precise the georeference is, permit flexible analyses across many resolutions. Because many records will be available at fine resolutions, this shift will enable many analyses to proceed at diverse resolutions, rather than being constrained at the outset.

Conclusions: the questions machine

The trend over the past decade for biodiversity data has been for different domains of data to become larger, better organized and more accessible (Edwards, 2008), at this point having resulted in on the order of 200 million occurrence records and almost 2 million names being organized and available online. We see some increased capacity to document and annotate data, thus increasing their fitness for use in informatics applications, and significant progress with designing (if not applying and using!) the stable identifiers that are necessary to enable linking across different domains. These advances, particularly if adopted widely and implemented soon, will improve greatly the global capacity to 'operate' jointly across domains.

The question, however, remains: *to what end?* Biodiversity informatics is not simply about provisioning data, however high the quality. Informatics subdisciplines can become central to the larger disciplines they serve by providing not simply old data, but by also producing new information and knowledge. Informatics disciplines help their communities understand new data more quickly as they are generated by leveraging existing data and tools. In biodiversity research, we believe that this potential exists, but despite the promise, we have not yet seen the delivery. What is missing? Our opinion is that what is missing are key conceptual and algorithmic approaches that link data to tools in order to generate new knowledge and answer key questions of interest to the biodiversity research community.

A very tangible example that is already available is simple visualization of species' known distributions in geographic dimensions. Geographic coordinates enable a painfully simple, but extremely useful, form of integration between biological and geospatial domains, and these visualizations (i.e. range maps, extent of occurrence maps, niche-model-derived range estimates) have become extremely popular (if not indispensable) in the biodiversity research and management community. Further integration of these data sources into our best estimates of species' distributions are possible, and such products would ultimately provide essential information, at relatively fine resolutions, about the arrangement of biodiversity across the planet. This key information product is nascent because the focus has been on raw data, rather than building conceptual approaches, formalized algorithmically, deployed across the vast set of data available and then served back to the community for use and annotation. These new information products themselves allow new questions to be asked, and it is the reciprocal building of knowledge that ultimately makes informatics disciplines into 'questions machines'. For example, a complete set of best estimates of species' geographic distributions would provide the information to ask new and interesting questions regarding processes that generate these distributions, which may be answerable by linkages to other data sources such as phenotypic and genetic data, or through species interactions databases.

By developing other such stable and known-quality links among data domains, other questions will be explored. Obvious link elements are scientific names, geographic coordinates, gene sequences, and eventually ontologies, which formalize entire classes of concepts and their relations. Scientific workflow software (Ludascher *et al.*, 2006), when more widely utilized, will become an analogue of GIS and enhance by orders of magnitude the capacity to establish standardized processes of gathering, linking, displaying and analysing different data domains. These methods will constitute veritable 'question machines', capable of helping a researcher to explore questions that previously would have taken perhaps years for lack of efficient access to data and tools, as our examples above illustrate.

However, even if such question machines become available very soon, true integration of biodiversity data cannot be achieved without theoretically deeper models and hypotheses about the processes by which organisms evolve and interact with others in a dynamic environment across spatiotemporal scales (Rangel *et al.*, 2007). It is not necessarily the case that new theories are needed, but rather that the databases and analytical tools should respond to such questions and fit within the conceptual frameworks. The level of integration envisioned in our future biodiversity informatics requires broad integration across domains, but in particular integration within conceptual frameworks set out by ecology, evolutionary biology and other overarching conceptual frameworks.

For a long time, biology has looked at the paradigms of physics for models of theoretical thinking, and conceptual integration often was made a synonym with equations, preferably derived from axiomatic first-principles (Martinez del Rio, 2008). However, biological objects are complex, history matters, and local details change in space and time in rather significant ways. True conceptual integration in biodiversity science should take these features as premises, rather than as nuisance, or as exceptions in elegant (but hopelessly simplified) theories. Logical representation of webs of concepts with the underlying very large bodies of actual data included as an organic part of the 'theories' may constitute an alternative conceptual path for biodiversity science. Our hope is that the initial steps forward in terms of improved data availability and analytical software availability will provide powerful incentives to move the field boldly in this direction.

References

Ackerly, D.D. 2003. Community assembly, niche conservatism, and adaptive evolution in changing environments. *International Journal of Plant Sciences* **164**, S165–S184.

Araújo, M.B. & New, M. 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* **22**, 42–47.

Ausubel, J.H. 2009. A botanical macroscope. *Proceedings of the National Academy of Sciences USA* **106**, 12569–12570.

Butler, D. 2006. Mashups mix data into global service. *Nature* **439**, 6–7.

Carstens, B.C. & Richards, C.L. 2007. Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution* **61**, 1439–1454.

Cracraft, J. 2002. The seven great questions of systematic biology: An essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanical Garden* **89**, 127–144.

Croizat, L. 1962. *Space, Time, Form: The Biological Synthesis*. Published by the author, Caracas, Venezuela.

Edwards, J.L. 2008. Research and societal benefits of the Global Biodiversity Information Facility. *BioScience* **54**, 486–487.

Fitzpatrick, B.M., Fordyce, J.A. & Gavrilets, S. 2009. Pattern, process and geographic modes of speciation. *Journal of Evolutionary Biology* **22**, 2342–2347.

Gotelli, N.J., Anderson, M.J., Arita, H.T., Chao, A., Colwell, R.K., Connolly, S.R., Currie, D.J., Dunn, R.R., Graves, G.R., Green, J.L., Grytnes, J.-A., Jiang, Y.-H., Jetz, W., Lyons, K., McCain, C.M., Magurran, A.E., Rahbek, C., Rangel, T.F.L.V.B., Soberón, J., Webb, C.O. & Willig, M.R. 2009. Patterns and causes of species richness: A general simulation model for macroecology. *Ecology Letters* **12**, 873–886.

Gotmark, F. & Nilsson, C. 1992. Criteria used for protection of natural areas in Sweden 1909–1986. *Conservation Biology* **6**, 220–231.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* **19**, 497–503.

Grayman, W.M., Males, R.M., Gates, W.E. & Hadder, A.W. 1975. Land-based modeling system for water quality management studies. *Journal of the Hydraulics Division, ASCE* **101**, 567–580.

Guralnick, R. & Hill, A. 2009. Biodiversity informatics: Automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* **25**, 421–428.

Hamid, J.S., Hu, P., Roslin, N.M., Ling, V., Greenwood, C.M.T. & Beyene, J. 2009. Data integration in genetics and genomics: Methods and challenges. *Human Genomics and Proteomics* **2009**, 869093.

Hey, T., Tansley, S. & Tolle, K. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, Redmond, Washington.

Hickerson, M., Carstens, B., Cavender-Bares, J., Crandall, K., Graham, C., Johnson, J., Rissler, L., Victoriano, P. & Yoder, A. 2009. Phylogeography's past, present, and future: 10 years after Avise 2000. *Molecular Phylogenetics and Evolution* **54**, 291–301.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965–1978.

Jackson, S.T. & Overpeck, J.T. 2000. Responses of plant populations and communities to environmental changes of the Late Quaternary. *Palaeobiology* **26** (supplement), 194–220.

Janies, D., Hill, A., Guralnick, R., Habib, F., Waltari, E. & Wheeler, W. 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Systematic Biology* **56**, 321–329.

Jetz, W. & Rahbek, C. 2001. Geometric constraints explain much of the species richness pattern in African birds. *Proceedings of the National Academy of Sciences USA* **98**, 5661–5666.

Kearney, M. & Porter, W.P. 2004. Mapping the fundamental niche: Physiology, climate, and the distribution of a nocturnal lizard. *Ecology* **85**, 3119–3131.

Kim, J.H. 2002. Bioinformatics and genomic medicine. *Genetics in Medicine* **4**, 62S–65S.

Knowles, L.L., Carstens, B.C. & Keat, M.L. 2007. Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Current Biology* **17**, 940–946.

Koleff, P., Tambutti, M., March, I., Esquivel, R., Cantú, C. & Lira-Noriega, A. 2009. Identificación de prioridades y análisis de vacíos para la conservación de la biodiversidad de México. In: Dirzo, R., González, R. & March, I., Eds., *Capital Natural de México. Estado de Conservación y Tendencias de Cambio*. Comision Nacional para el Conocimiento y Uso de la Biodiversidad, Mexico, D.F., pp. 651–718.

Koning, D., Sarkar, I. & Moritz, T. 2005. TaxonGrab: Extracting taxonomic names from text. *Biodiversity Informatics* **2**, 79–82.

Koslow, S. & Hirsch, M. 2004. Celebrating a decade of neuroscience databases. *Neuroinformatics* **2**, 267–269.

Kozak, K.H., Wiens, J.J. & Pfennig, D. 2009. Does niche conservatism promote speciation? A case study in North American salamanders. *Evolution* **60**, 2604–2621.

Losos, J.B. & Glor, R.E. 2003. Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology and Evolution* **18**, 220–227.

Ludascher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J. & Zhao, Y. 2006. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience* **18**, 1039–1065.

Lynch, J.D. 1989. The gauge of speciation: On the frequency of modes of speciation. In: Otte, D. & Endler, J.A., Eds., *Speciation and its Consequences*. Sinauer Associates, Sunderland, MA, pp. 527–553.

Manso, M.-Á. & Wachowicz, M. 2009. GIS design: A review of current issues in interoperability. *Geography Compass* **3**, 1105–1124.

Martínez-Meyer, E., Peterson, A.T. & Hargrove, W.W. 2004. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography* **13**, 305–314.

Martinez Del Rio, C. 2008. Metabolic theory or metabolic models? *Trends in Ecology and Evolution* **25**, 256–260.

Nelson, G.O. & Platnick, N.I. 1981. *Systematics and Biogeography: Cladistics and Vicariance*. Columbia University Press, New York.

Page, R.D.M. 2008. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics* **9**, 345–354.

Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sanchez-Cordero, V., Soberon, J., Buddemeier, R.H. & Stockwell, D.R.B. 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature* **416**, 626–629.

Peterson, A.T., Soberón, J. & Sánchez-Cordero, V. 1999. Conservatism of ecological niches in evolutionary time. *Science* **285**, 1265–1267.

Porter, W., Sabo, J., Tracy, C., Reichman, O. & Ramankutty, N. 2002. Physiology on a landscape scale: Plant-animal interactions. *Integrative and Comparative Biology* **42**, 431–453

Rahbek, C., Gotelli, N.J., Colwell, R.K., Entsminger, G.L., Rangel, T.F.L.V.B. & Graves, G.R. 2007. Predicting continental-scale patterns of bird species richness with spatially explicit models. *Proceedings of the Royal Society B* **274**, 165–174.

Rangel, T.F., Diniz-Filho, J.A. & Colwell, R. 2007. Species-richness and evolutionary niche dynamics: A spatial pattern-oriented simulation experiment. *American Naturalist* **170**, 602–616.

Ricklefs, R.E. & Schluter, D. 1993. *Species Diversity in Ecological Communities: Historical and Geographical Perspectives*. University of Chicago Press, Chicago.

Ridgely, R.S., Allnutt, T.F., Brooks, T., Mcnicol, D.K., Mehlman, D.W., Young, B.E. & Zook, J.R. 2005. *Digital Distribution Maps of the Birds of the Western Hemisphere, version 2.1*. NatureServe, Arlington, Virginia.

Roos, D.S. 2001. Bioinformatics—Trying to swim in a sea of data. *Science* **291**, 1260–1261.

Sarkar, S., Pressey, R.L., Faith, D.P., Margules, C.R., Fuller, T., Stoms, D.M., Moffett, A., Wilson, K.A., Williams, K.J., Williams, P.H. & Andelman, S. 2006. Biodiversity conservation planning tools: Present status and challenges for the future. *Annual Review of Environment and Resources* **31**, 123–159.

Sautter, G., Boehm, K. & Agosti, D. 2006. A combining approach to find all taxon names (FAT). *Biodiversity Informatics* **3**, 46–58.

Schalk, P.H. 1998. Management of marine natural resources through biodiversity informatics. *Marine Policy* **22**, 269–280.

Soberón, J. & Nakamura, M. 2009. Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences USA* **106**, 19644–19650.

Soberón, J. & Peterson, A.T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics* **2**, 1–10.

Sonnhammer, E.L.L. 2004. Genome informatics: Taming the avalanche of genomic data. *Genome Research* **6**, 301.

Stein, B.R. & Wieczorek, J. 2004. Mammals of the world: MaNIS as an example of data integration in a distributed network environment. *Biodiversity Informatics* **1**, 14–22.

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., Ferreira De Siqueira, M., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A.S., Midgely, G.E., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L. & Williams, S.E. 2004. Extinction risk from climate change. *Nature* **427**, 145–148.

Vogler, A.P. & Monaghan, M.T. 2006. Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research* **45**, 1–10.

Waltari, E., Perkins, S., Hijmans, R., Peterson, A.T., Nyári, Á. & Guralnick, R. 2007. Locating Pleistocene refugia: Comparing phylogeographic and ecological niche model predictions. *PLoS ONE* **2**, e563.

Wells, P.V. 1983. Palaeobiogeography of montane islands in the Great Basin since the last glaciopluvial. *Ecological Monographs* **53**, 341–382.

Williams, J.W. & Jackson, S.T. 2007. Novel climates, no-analogue communities, and ecological surprises. *Frontiers in Ecology and the Environment* **5**, 475–482.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. & Culham, A. 2007. How global is the Global Biodiversity Information Facility? *PLoS ONE* **2**, e1124.