The encyclopedia of life

Edward O. Wilson

Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138-2902, USA

Comparative biology, crossing the digital divide, has begun a still largely unheralded revolution: the exploration and analysis of biodiversity at a vastly accelerated pace. Its momentum will return systematics from its long sojourn at the margin and back into the mainstream of science. Its principal achievement will be a single-portal electronic encyclopedia of life.

Imagine an electronic page for each species of organism on Earth, available everywhere by single access on command. The page contains the scientific name of the species, a pictorial or genomic presentation of the primary type specimen on which its name is based, and a summary of its diagnostic traits. The page opens out directly or by linking to other data bases, such as ARKive, Ecoport, GenBank and MORPHOBANK. It comprises a summary of everything known about the species' genome, proteome, geographical distribution, phylogenetic position, habitat, ecological relationships and, not least, its practical importance for humanity.

The page is indefinitely expansible. Its contents are continuously peer reviewed and updated with new information. All the pages together form an encyclopedia, the content of which is the totality of comparative biology.

The rationale

There are compelling reasons to build such an all-species encyclopedia. Not least is the heuristic power for biology as a whole. As the census of species on Earth comes ever closer to completion, and as their individual pages fill out to address all levels of biological organization from gene to ecosystem, new classes of phenomena will come to light at an accelerating rate. Their importance cannot be imagined from our present meagre knowledge about the biosphere and the species comprising it. Who can guess what the mycoplasmas, collembolans, tardigrades and other diverse and still largely unknown groups will teach us? As the species coverage grows, gaps in our biological knowledge will stand out like blank spaces on maps. They will become destinations toward which researchers will gravitate.

For the first time, the biotas of entire ecosystems can be censused in full. Unknown microorganisms and the smallest invertebrates, which still comprise most species yet lack even a name, will be revealed. Only with such encyclopedic knowledge can ecology mature as a science and acquire predictive power species by species, and from those, ecosystem by ecosystem.

As one result, the human impact on the living environment could be assessed in far more reliable detail than is now possible. Today, for example, we base estimates of species extinction on data from a scattering of taxonomically best known groups, including the flowering plants, land and freshwater vertebrates, and a few invertebrates, such as butterflies and mollusks. These taxa contain only about a quarter of the known species on Earth, and almost certainly a much smaller fraction of those still unknown. Tomorrow, other invertebrates, including insects and nematodes, as well as fungi and nearly all microorganisms, together comprising most species on Earth, as well as essential pathways of the energy and materials cycles, can also be assessed.

The all-species encyclopedia will serve human welfare in more immediately practical ways. The discovery of wild plant species adaptable for agriculture, new genes for enhancement of crop productivity, and new classes of pharmaceuticals can be accelerated. The outbreak of pathogens and harmful plant and animal invasives will be better anticipated and halted. Never again, with fuller knowledge of such extent, need we overlook so many golden opportunities in the living world around us, or be so often surprised by the sudden appearance of destructive aliens that spring from it.

An all-species encyclopedia of life is logically inevitable if for no other reason that the consolidation of biological knowledge is urgently overdue. In its earliest stages, already emerging, it forms a matrix within which comparative studies are rapidly organized. The process will accelerate as traditional taxonomic procedures, still mostly dependent on repeated examinations of type specimens and print literature, are replaced by high-resolution digital photography, nucleic acid sequencing and internet publication. With further documentation organized into the species pages, new lines of research will open at a quickening pace. Model species for laboratory and field research can be more easily found – obedient to the principle that for every problem in biology, there exists a species ideal for its solution.

A growing, single-access species-structured encyclopedia will ease navigation through the immense biological data bases. Aided by computer search engines, patterns can be summoned whose detection would otherwise demand impracticable amounts of effort and time. Principles and theory can be built, deconstructed and rebuilt with an unprecedented power and transparency.

Ultimately, and at a deeper level, the all-species encyclopedia will, I believe, transform the very nature of biology, because biology is primarily a descriptive science. Although it depends upon a solid base of physics and chemistry for its functional explanations, and the theory of natural selection for its evolutionary explanations, it is

Corresponding author: Edward O. Wilson (ewilson@oeb.harvard.edu).

http://tree.trends.com 0169-5347/02/$ - see front matter PII: S0169-5347(02)00040-X

defined uniquely by the particularity of its elements. Each species is a small universe in itself, from its genetic code to its anatomy, behavior, life cycle and environmental role, a self-perpetuating system created during an almost unimaginably complicated evolutionary history. Each species merits careers of scientific study and celebration by historians and poets. Nothing of the kind can be said (at the risk of stating the obvious) for each proton or inorganic molecule.

The taxonomic foundation

Taxonomy, the scientific study and practice of classification, is the foundation to the all-species encyclopedia. However, it is still one of the most underfunded and weakly developed biological disciplines. Worldwide, as few as 6000 biologists work within it. Most people are surprised to learn that most of biodiversity is still entirely unknown. They assume that taxonomy all but wound down generations ago, so that today each new species discovered is a newsworthy event. The truth is that we do not know how many species of organisms exist on Earth even to the nearest order of magnitude. Those formally diagnosed and given latinized scientific names are thought to number somewhere between 1.5 and 1.8 million, with no exact accounting having yet been made from the taxonomic literature. Estimates of the full number, known plus unknown, vacillate wildly according to method. As summarized in the *Global Biodiversity Assessment* [1], they range from an improbable 3.6 million at the low end to an equally improbable 100 million or more at the high end. The commonest order-of-magnitude guess is ten million.

The smaller the organisms, the more poorly known the group to which it belongs. About 69 000 species of fungi have been distinguished and named, but as many as 1.6 million are thought to exist. Of the nematode worms, making up to four of every five animals on Earth (and, it is said, so abundant that if all solid matter on the surface of the planet were to disappear, its ghostly outline could still be seen in nematodes), ~15 000 species are known but millions more might await discovery. Nematodes in turn are dwarfed in diversity by the bacteria and archaeans, the black hole of biological systematics. Although only ~6000 have been formally recognized, approximately that many, almost all new to science, can be found in only a few grams of rich forest soil. Our ignorance of these microorganisms is epitomized by bacteria of the genus *Prochlorococcus*, arguably the most abundant organisms on the planet and responsible for a large part of the organic production of the ocean, yet unknown until 1988. *Prochlorococcus* cells float passively in open water at 70 000–200 000 ml^{-1}, multiplying with energy captured by sunlight. They eluded recognition so long because of their extremely small size. Representing a special group called picoplankton, they are much smaller than conventional bacteria and barely visible at the highest optical magnification.

Even the largest organisms await a full accounting. The global number of amphibian species has grown in the past 15 years by more than a third, from 4000 to 5400. The flowering plants, for centuries among the favorite targets of naturalists, could rise from the present 272 000 to over 300 000: each year ~2000 new species are added to the standard world list of the *International Plant Names Index* (http://www.ipni.org).

The biodiversity agenda

How best might the taxonomic foundation be laid? From 13 to 15 October, 2001, a 'summit' was held at Harvard University by leaders of organizations devoted to comprehensive taxonomic surveys on a global or continental scale. Their aim was to find a way to complete a world census in a foreseeable period of time. Included were the Africa Biodiversity Foundation (headquartered in Bulawayo, Zimbabwe), Census of Marine Life (New York, USA), the Global Biodiversity Information Facility (Copenhagen, Denmark), the Global Taxonomy Initiative of the Convention on Biological Diversity (New York), the Integrated Taxonomic Information System (Washington, DC, USA), and NatureServe (Arlington, USA). Also present were scientist representatives from major collections in North and Latin America, as well as experts in bioinformatics technology. The summit was hosted by the All Species Foundation, newly formed as a facilitator of the overall effort. Its aim is to provide a clearing-house for the frontline initiatives, to assist them in their funding initiatives and development of bioinformatics, to initiate new projects, and to monitor and report progress in the overall enterprise on a continuing basis.

The attendees of the all-species summit agreed that a complete or, more realistically, a nearly complete global biodiversity census is technically feasible within 25 years. The magnitude of the task can be visualized as follows: whereas 10% of species on Earth out of, say (at an educated guess) 10 million–20 million, have been diagnosed during the first 250 years, beginning with Carolus Linnaeus' *Systema Naturae* in the mid-1700s, it is proposed to complete the remaining 90% in one-tenth that time.

The idea of a complete global biodiversity census with a timeline and coordinated initiatives had first been proposed in 1992 [2]. By the mid-1990s, the importance of the new technologies of bioinformatics in descriptive biology had also become apparent [3]. In 2000, explicit proposals were put forth for a census timeline and practical bioinformatics in systematics research [4–8]. By 2002, the implications of the new initiatives were being explored by biologists in several disciplines [9–11], and it could be said quite fairly that a 'biodiversity commons' [12] had come into being within the 'bioinformatics nation' [13].

The full agenda of biodiversity exploration is now unfolding in three overlapping phases. The first is the Catalog of Life, aimed at the organization of information about existing species into an electronic global framework [11]. The Catalog was born of the collaborative efforts of Species 2000, a federation of data bases begun in 1994 by the International Union of Biological Sciences, and headquartered at the University of Reading, UK; the Integrated Taxonomic Information System, begun in 1995

through a partnership among interested agencies of the US Federal Government; the Global Taxonomy Initiative of the Convention on Biological Diversity, a worldwide effort spun from the 1992 Rio Earth Summit; and the Global Biodiversity Information Facility, begun by the Organization for Economic Cooperation and Development in 1996 and now headquartered as an independent operation in Copenhagen.

The second phase of the full biodiversity agenda is the accelerated discovery of life forms still unknown. This achievement, the anticipated moon shot of systematic biology, is envisioned as a future goal by the organizations loosely grouped under what Bisby *et al.* [11] have called the 'Catalog of Life' initiative, and as an immediate goal with a timeline by the All Species Foundation, headquartered in San Francisco, USA [6–8,10].

The final enterprise, the electronic Encyclopedia of Life, which is already being pressed here and there, will expand upon the growing base provided by the taxonomic Catalog of Life. Covering all biological levels, from genome to ecology, it will serve as the ultimate guide to biodiversity.

New technologies

Faith in a sprint to the finish of the global census is engendered by the more advanced revolutions ongoing in bioinformatics and genomics, which together offer the means to transform the traditional methods of taxonomy. The old methods, which still prevail, have been enormously labor intensive and time consuming. To complete a taxonomic analysis of a genus or higher order taxon requires examination of the primary types of each species, subspecies and variety, which are typically scattered among museums in North America and Europe, and often in other continents. The systematist must conduct lengthy tours to examine all these specimens, or else have them sent through by hand or mail, a risky step that not all curators are willing to take. The systematist must also have access to a wide array of books and journals, many of which are old and rare. As a result, the tradition of systematics since Linnaeus has been that of arcane expertise practiced by groups of specialists working on groups of organisms to which they have devoted their professional lives.

With the new technology, the 19th century culture of taxonomy has begun to be replaced. For the first time, type specimens can be illustrated by swiftly made high-resolution digital photographs, the anatomical detail and depth of field of which are beyond those seen in specimens viewed by light microscopy. The photographs can be published on the Internet. When all the primary types of a particular group, say weevils of the family Curculionidae or grasses of the family Gramineae, are digitally photographed and online, they can be accessed immediately by anyone anywhere. When the original diagnoses from print literature are added, experts can proceed with revisions at a speed and an economy vastly greater than enjoyed in the predigital era. In one step, the practice of taxonomy is globalized and democratized and, in a sense, the type specimens are repatriated to their country of origin.

One such program already completed is the 'virtual herbarium' of the New York Botanical Garden. Almost its entire collection of type specimens of some vascular plants, representing 90 000 species, is now finished. Similar initiatives are underway in the insect collections of the Academy of Natural Sciences in Philadelphia, USA and Harvard University's Museum of Comparative Zoology. With more such projects completed, collection by collection around the world, the global iconography will come together like pieces fitted into a mosaic. The result will be the requisite foundation for a swift exploration of biodiversity on Earth and the accompanying growth of the all-species encyclopedia.

Key challenges

Construction of the complete taxonomic base will not, however, be just a smooth compilation of species. The magnitude of biodiversity and the tangle of evolutionary processes that generated it still present formidable problems. First in line is the difficulty of classifying microorganisms and many of the smallest, soft-bodied invertebrates, most of the species of which can be reliably separated only by molecular diagnosis. The difficulty has put all-species inventories out of reach in the past. However, its solution appears close at hand, thanks to the rapid advances occurring in genomics. Already, for example, tens of thousands of species from the major domains of organisms have been at least partially sequenced for small subunit rRNA genes. By April 2002, the last date for which I have seen an accounting, the genomes of no fewer than 61 species of bacteria had been completely sequenced. As the process accelerates, and the cost per base pair continues to drop, genomic data will become standard for taxonomy, as well as for phylogenetic reconstruction, across all groups of organisms.

A second barrier to the all-species inventory is the incongruence of the species concept between major groups. The classic definition of the species in sexually reproducing organisms is a closed gene pool – a population of individuals that are capable of freely interbreeding under natural conditions. This criterion works reasonably well for most animals and plants, but creates difficulties in some plant groups in which hybridization is extensive but short of total. And it fails logically, of course, in the many populations that lack sexual reproduction. The value of the classic definition of reproductive isolation is still unknown in the bulk of microorganisms, where species might have to be delineated arbitrarily by a cutoff percentage of base pairs shared by populations or some other genetic criterion.

The species problem cannot be settled in advance by any formula or legislation. It will probably be broken only as the all-species initiative evolves, illuminating the particularities of species-level variation from one phylogenetic group of organisms to another. As this knowledge grows, the difficulty of defining species will metamorphose into deeper studies of how species-level diversity arises, group by group. Meanwhile, the process of censusing can and should proceed with the best tools and species concepts at

hand. Resolution of the species problem will be one of its most important results.

The problems inherent in bioinformatics are also formidable. As electronic search engines are developed, they must be made interoperable within and between phylogenetic groups. They must have quality control, exercised most probably by publication committees comparable to boards of editors of journals. They need to be created, as in the case of GenBank, to provide free public access. In joining the bioinformatics nation, taxonomists and encyclopedists need to address and overcome the growing problem of information overload already bedeviling those managing DNA microarray analyses, airline schedules and bank accounts. And finally, with current floppy disks starting to lose data within a decade and even optical disks in less than a century, improvement in longevity and format transfer methods will be a priority in the technologies adopted.

These obstacles are daunting, but they are of a technical nature eminently vulnerable to human ingenuity. To overcome them, and thereby complete the great Linnaean enterprise, creating the base of the all-species encyclopedia, will secure the rightful place of comparative biology within mainstream science.

References

1 Heywood, V.H. and Watson, R.T. (1995) *Global Biodiversity Assessment*, Cambridge University Press
2 Raven, P.H. and Wilson, E.O. (1992) A fifty-year plan for biodiversity surveys. *Science* 258, 1099–1100
3 Edwards, M. and Morse, D.R. (1995) The potential for computer-aided identification in biodiversity research. *Trends Ecol. Evol.* 10, 153–158
4 Wilson, E.O. (2000) A global biodiversity map. *Science* 289, 2279
5 Wilson, E.O. (2000) On the future of conservation biology. *Conserv. Biol.* 14, 1–3
6 Kelly, K. (2000) All species inventory: a call for the discovery of all life-forms on Earth. *Whole Earth* Fall, 4–9
7 Warshall, P. (2000) Bioinformatics: the master list and virtual museum. *Whole Earth* Fall, 50
8 Lawler, A. (2001) Up for the count? *Science* 294, 769–770
9 Godfray, C.J. (2002) Challenges for taxonomy. *Nature* 417, 17–19
10 Gerwin, V. (2002) All living things, online. *Nature* 418, 362–363
11 Bisby, F.A. *et al.* (2002) Taxonomy, at the click of a mouse. *Nature* 418, 367
12 Moritz, T. (2002) Building the biodiversity commons. *D-Lib Magazine* 8 http://www.dlib.org/dlib/june02/moritz/06moritz.html
13 Stein, L. (2002) Creating a bioinformatics nation. *Nature* 417, 119–120