

# DRAFT FRAMEWORK FOR LAUNCHING THE ENCYCLOPEDIA OF LIFE

## DISTRIBUTION DRAFT FOR COMMENT

“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 databases, 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 perceived practical importance for humanity.

“The page is indefinitely expandable. 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,”

E. O. Wilson, “The encyclopedia of life,” *Trends in Ecology and Evolution*, 18 (2003)

### Cornerstone Institutions and Principal Investigators:

Smithsonian Institution ([SamperC@si.edu](mailto:SamperC@si.edu))

Marine Biological Laboratory ([gborisy@mbl.edu](mailto:gborisy@mbl.edu))

Biodiversity Heritage Library consortium ([garnettT@si.edu](mailto:garnettT@si.edu))

Field Museum of Natural History ([jmccarter@fieldmuseum.org](mailto:jmccarter@fieldmuseum.org))

Harvard University ([hanken@oeb.harvard.edu](mailto:hanken@oeb.harvard.edu))

### Honorary Chair:

Edward O. Wilson

Museum of Comparative Zoology

Harvard University

## SUMMARY

The idea for the Encyclopedia of Life (EOL) is deceptively simple: construct a web site for each of the approximately 1.8 million species now known to be present on Earth, and make them all accessible through a single portal. Each site will consist of several web pages. The entry page, which the user may configure, will present information about the species suitable for the general public and contain, for example, the scientific and common names (in several languages) by which the species is known and has been known in the past, an illustration, a range map, habitat and natural history, conservation status and human uses, and how to identify it. This entry page is much like a good field guide.

The power of the EOL, however, lies in the fact that “species name” is a field in almost every biological database, and so can be used to organize and link to virtually the totality of biological knowledge. From the entry page, itself dynamically synthesized, users will be able to jump to a wide variety of more specialized pages, including ones for taxonomists, molecular biologists, geneticists, those working on genomics (and all the other “-omics”), physiologists, teachers, horticulturists, phylogeneticists, conservationists, specimen and observational databases, etc., etc. As long as the species name is a field in another database, its data will be able to be linked into the species site. Biodiversity literature will be available, too, through the EOL’s partnership with the Biodiversity Heritage Library (BHL) consortium. As part of the EOL, the BHL will digitize the world’s biodiversity literature, and EOL will work with BHL to develop navigational tools to easily move around that literature and between the literature and the species pages.

Demand for biodiversity information by the scientific community, natural resource managers, the education sector and the public at large is growing. We believe the time for the EOL is right because there has been important progress in recent years in making information available and the technology has advanced to a point where this knowledge can be linked together in an efficient and cost-effective way. *Over the next five years, the EOL aims to (a) generate a million species pages, each one of which will be verified by an expert, (b) digitize a large portion of biodiversity literature, (c) generate educational materials for students, schools and universities and (d) use the EOL resource to generate new synthetic knowledge about the world’s biodiversity.*

This project brings together several of the world’s leading natural history institutions: the Smithsonian Institution, the Field Museum, Harvard University, the Marine Biological Laboratory and the Biodiversity Heritage Library consortium. Other institutions from across the United States and around the world will be invited as collaborators, and links have already been established with many relevant efforts already underway.

The cost for the first five-year phase of the EOL is estimated to be approximately \$40 million, and it is anticipated that sufficient funds to begin implementing the project will be available by the middle of 2007. The EOL will be led by a Board of Directors of leaders of the cornerstone institutions, an Advisory Committee of distinguished individuals to orient and advocate the project, an Institutional Council of other partner organizations and programs key to its realization, and a Secretariat to conduct its core activities. In order gain buy-in and involvement from key partners in the US and worldwide , the Board of Directors will be convening a series of planning meetings during the first half of 2007. These meetings will provide advice in preparing detailed plans and milestones for the major components of the project, including the mashup software for producing the provisional species sites; a mechanism for involving the community of experts in reviewing and certifying the sites; scanning and digitization of the literature: education and outreach activities; and a synthesis center to use the resources provided by the EOL to spur new knowledge.

## A. CONCEPT

In 1979, Joël de Rosnay published a fascinating book called “The Macroscope.” De Rosnay, a molecular biologist, systems theorist and futurologist, posited that just as microscopes let us see things that are too small for the naked eye to discern, we need to develop macroscopes to discern things that are too large for us to see otherwise. This metaphor has been applied to economics, astronomy and several other fields. With the Encyclopedia of Life (EOL), we are now proposing to construct what will, in essence, become a macroscope for biodiversity and an entry point into virtually all of biological knowledge.

The Encyclopedia of Life (EOL) is an ambitious, even audacious project to organize and make available via the Internet virtually all information about life present on Earth. At its heart lies a series of web sites—one for each of the approximately 1.8 million known species—that provide the entry points to this vast array of knowledge. The entry-point for each site will be a species page suitable for the general public, but with several linked pages aimed at more specialized users. The sites will sparkle with text and images that will be enticing to everyone, as well as providing deep links to specialized knowledge.

The EOL derives its power from the fact that “species name” is the only field common to virtually all biological databases. Thus, using this field it can link together widely disparate information, ranging from molecular to ecosystem studies, as well as the literature about them, while at the same time being accessible to, and usable by, the general public.

The EOL will dynamically synthesize biodiversity knowledge about all known species, including their taxonomy, geographic distribution, collections, genetics, evolutionary history, morphology, behavior, ecological relationships, and importance for human well being, and distribute this information through the Internet. It will serve as a primary resource for a wide audience that

includes scientists, natural resource managers, conservationists, teachers, and students around the world. The EOL's encompassing scope and innovation will have a major global impact in facilitating biodiversity research, conservation, and education.

The need for the EOL is clear as the impact of human activities on biodiversity increases and threatens many species. Population growth, global climate change and other environmental changes heighten the urgency with which we must discover, understand and protect biodiversity.

Many countries, especially those in the developing world, include biodiversity as an important element in their economic and social development. The EOL will greatly facilitate access to biodiversity information in the countries where most of the planet's biodiversity is found and where it must be sustainably managed.

The EOL is also timely because rapid development of information technology is opening doors to huge potential for new and innovative approaches. Recent informatics technologies, such as *wikis* and *aggregation technology*, make the rapid assembly of the EOL possible. The EOL will use aggregation technology to create hundreds of thousands of provisional web pages for species, and then employ a peer-reviewed wiki approach to allow biodiversity experts to evaluate and upgrade the pages. Writing an Encyclopedia of Life the traditional way would have required many years just to identify experts to write each article and then decades for the experts to deliver the articles. The EOL approach frees experts of the burden of drafting, makes this the job of software, and optimizes the use of precious expert time. This precious time could focus on quality control, generating new content for little-known species, and other non-routine functions.

In addition, the recent capability to scan published materials at a rapid rate and to mark-up the digitized pages means that the EOL will be able to link directly to a wealth of published material. Through the involvement of the Biodiversity Heritage Library consortium, EOL users will

be able to seamlessly access millions of pages of biodiversity literature. All this literature will be available via open-source principles, which will allow users to develop value-added applications to mine the literature in ways heretofore impossible.

The EOL will be a locus for integrating many different kinds of stakeholders. Libraries, natural history museums, universities and research institutes, national and international agencies concerned with environment, agriculture, and health, non-governmental organizations, commercial information providers, the education sector, and the general public are among the stakeholders. The EOL offers the chance to integrate their interests for public benefit. No comparable initiative exists on this scale, and no existing forum or organization could encompass the task.

The EOL is of a similar magnitude to the Sloan Digital Sky Survey. Just as that project has led to many advances in our understanding of the heavens, the EOL is expected to have a catalytic effect on studies of biodiversity. The pioneering biologist E. O. Wilson said it well: “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.”

For further background information about the EOL, including its rationale and justification, see Appendix 2.

## **B. STRUCTURE**

The EOL is intended to be a bridge between science and society and between scientists and citizens, as well as a research environment for scientists. The ultimate goal of the EOL is to build a consumer-driven product that provides the most authoritative information on all species and the means to add, mine and analyze the information. To accomplish this goal, the EOL is composed of the following five core elements.

**Species sites.** The central component of a species site will be the entry-level species page for the general public. This will contain a standardized series of components, which will vary for different kinds of organisms, but will be configurable by individual users so that they will see the information they want to see (“my EOL”). Common to all entry-level species pages will be the scientific and common names (in several languages) of the species; a species description (comprehensible by the general public); picture(s) and/or illustrations of the species; habitat information; a range map to show where the species lives; instructions on how to identify the members of the species; information on its natural history, conservation status and human uses (where these are known); and some bibliographic information. The exact components of the entry-level species page will be determined by a structured needs analysis of the general users, and by the editorial and design committee and by experts involved with the kind of organism (you can’t have the same blocks of information, for example, for a gray squirrel and a paramecium). However, as much as possible the entry-level pages will have a common layout, look and feel.

Linked to the entry-level page will be a varying number of more specialized pages covering different domains of knowledge—taxonomy, molecular biology, physiology, conservation, phylogeny, education, literature, specimen and observational data, biogeography, etc. As data sets become available in these and other areas, they will be linked to the species sites, allowing the EOL to be extensible and sustainable.

Several international web-based projects to organize biodiversity knowledge in various domains—such as the Global Biodiversity Information Facility ([www.gbif.org](http://www.gbif.org)), FishBase ([www.fishbase.org](http://www.fishbase.org)), Species2000 ([www.sp2000.org](http://www.sp2000.org)), and the International Plant Name Index ([www.ipni.org](http://www.ipni.org))—are now functioning, and EOL will partner with them, thus leveraging their content in ways that will be synergistic for all the partners. We estimate that more than 100,000

species pages are already represented on the Internet. However, every group (ants, fish, ferns) has pages with different fields and designs. While many pages have rich content and imaginative design, the Babel makes life hard for the user. Moreover, few are using aggregation technology, so the pages are static and may become quickly dated. And few have thought about how the pages might be transformed for access through personal digital assistants or cell phones in the field. The EOL can set high standards and key norms, including for interoperability and linking of resources. Building on much excellent work accomplished in biodiversity informatics during the past 20 years, the EOL offers a chance to multiply content and convenience by orders of magnitude.

The EOL is species-centric, with its eyes firmly focused on developing the species sites, and the rapid creation of these sites at a high level of quality is the major goal of the enterprise. However, because many users will want to enter the Encyclopedia at other levels of classification (e.g. mammals, whales, Asteraceae), we do need to be prepared to meet needs in this regard, and the EOL will thus contain entry points, descriptions and linking mechanisms to deal with these higher levels.

EOL must show some results and value quickly. Therefore, we would seek first to work with existing compendia of species pages, such as FishBase, AmphibiaWeb, and North American Mammals. Vertebrates and angiosperms will be covered early, which has the plus that these are largely organisms the general public and educators will find of high value. More difficult groups will be tackled as the scientific societies and experts indicate their readiness to deal with them.

**Informatics.** One key to the success of the EOL is a new cyberinfrastructure that will engage numerous initiatives that share a common interest in the collection and dissemination of expert information about organisms. Several types of software will be needed. Software for wikis, other web services and aggregation technologies, and for running large portals is now widely available

and will be adapted and fine-tuned for the EOL. Already in development are navigational tools for searching and indexing the species sites and tools for marking-up of the digitized literature, including automated mechanisms for identifying scientific names.

The cyberinfrastructure has three important dimensions. The first is a layered organizational architecture that brings together any information on any organism and assembles that information within biologically meaningful contexts. Secondly, “taxonomic intelligence” will be used to overcome traditional problems of bringing together distributed data on the Internet. Thirdly, EOL will use innovative aggregation technology (Mashups) to draw together information from multiple sources and reassemble the information with sufficient flexibility that the resulting web pages can meet the needs of diverse audiences.

For more information, see Appendix 3, EOL Technical Vision.

**Literature.** The Biodiversity Heritage Library (BHL) project ([www.bhl.si.edu](http://www.bhl.si.edu)) is a consortium of ten of the world’s largest natural history libraries (American Museum of Natural History, Field Museum, Harvard University Botany Libraries, Ernst Mayr Library of the Museum of Comparative Zoology, Library of the Marine Biological Laboratory and Woods Hole Oceanographic Institution, Missouri Botanical Garden, Natural History Museum London, New York Botanical Garden, and Smithsonian Institution Libraries). The BHL is preparing to digitize the contents of the member libraries, thus making available virtually the entire store of published human knowledge about biodiversity. Linking the literature information to the data available through the EOL species sites provides great depth that will make the EOL a much greater scientific and popular resource. It will also make this information available to scientists and the general public in developing countries. It is a vexing contrast that most of the world’s biodiversity is found in developing countries, while most of the information about that biodiversity is locked up in the literature and databases of the

developed countries. EOL will contribute to “repatriating” that knowledge back to the countries of origin.

The BHL estimates that the total published biodiversity literature consists of 280 to 320 million pages, of which 170-190 million pages form the “core” literature, and about 80 million a base that includes some information on most species. It is developing a “union catalogue” of the holdings of the ten libraries that make up the consortium and is developing rational and efficient strategies to digitize the core literature. In addition, some institutions and organizations have already digitized portions of the biodiversity literature (e.g., several natural history museums have digitized their own publications), totaling perhaps 1 million pages, and it may be feasible to incorporate these into the project. The digitizing process will be organized around regional scanning sites, using technology and processes already developed by the Internet Archive ([www.archive.org/index.php](http://www.archive.org/index.php)).

The BHL anticipates opening one or two scanning sites in 2007, each of which would be capable of scanning up to one million pages a month. Software is under development to bring “taxonomic intelligence” to the scanned pages (for further details, see “Indexing content from the Biodiversity Heritage Library” in Appendix 3).

**Education and Outreach.** The EOL will present a treasure-trove of information for use in educational environments, ranging from grades K-12 through graduate levels, as well as a wealth of possibilities for informal education. EOL will partner with education-oriented institutions and organizations and with relevant citizen scientists (e.g., horticulturists, birdwatchers) to develop curricula, lesson plans, and tools for educational uses of the EOL and for outreach to the general public. Models will also be developed for using the EOL in citizen science, where students and the general public can be involved in gathering information and asking questions.

The Education and Outreach component of EOL will also identify the most efficient mechanisms for outreach to a variety of audiences and determine the costs and impacts for each audience.

**Synthesis center.** Experience has shown that every time a group develops a large amount of scientific data, the very existence of the datasets enables new scientific questions and hypotheses. Examples of this phenomenon can be seen at the Sloan Digital Sky Survey ([www.sdss.org/background/](http://www.sdss.org/background/)) and the National Center for Ecological Analysis and Synthesis ([www.nceas.ucsb.edu](http://www.nceas.ucsb.edu)).

The EOL will quickly become a nexus for integration of biodiversity information from many fields. New ways of analyzing and combining large data sets will be required to make full synergistic use of EOL's broad and varied resources.

The EOL synthesis center will facilitate scientific analysis and educational utility of the base data sets collated by the EOL. It will play three major roles: (1) aiding scientific working groups developing tools and novel uses for the EOL; (2) facilitating cross-disciplinary involvement of the scientific community in the EOL effort; and (3) generating new scientific questions and hypotheses that will demonstrate the power of the EOL.

The center will serve as the locus for bringing together groups of scientists to examine how to catalyze such innovations using the information gathered in the species sites. Suitable topics for such approaches might include, for example, phylogenetic and biogeographic analyses, or studies on coevolution of taxonomic groups.

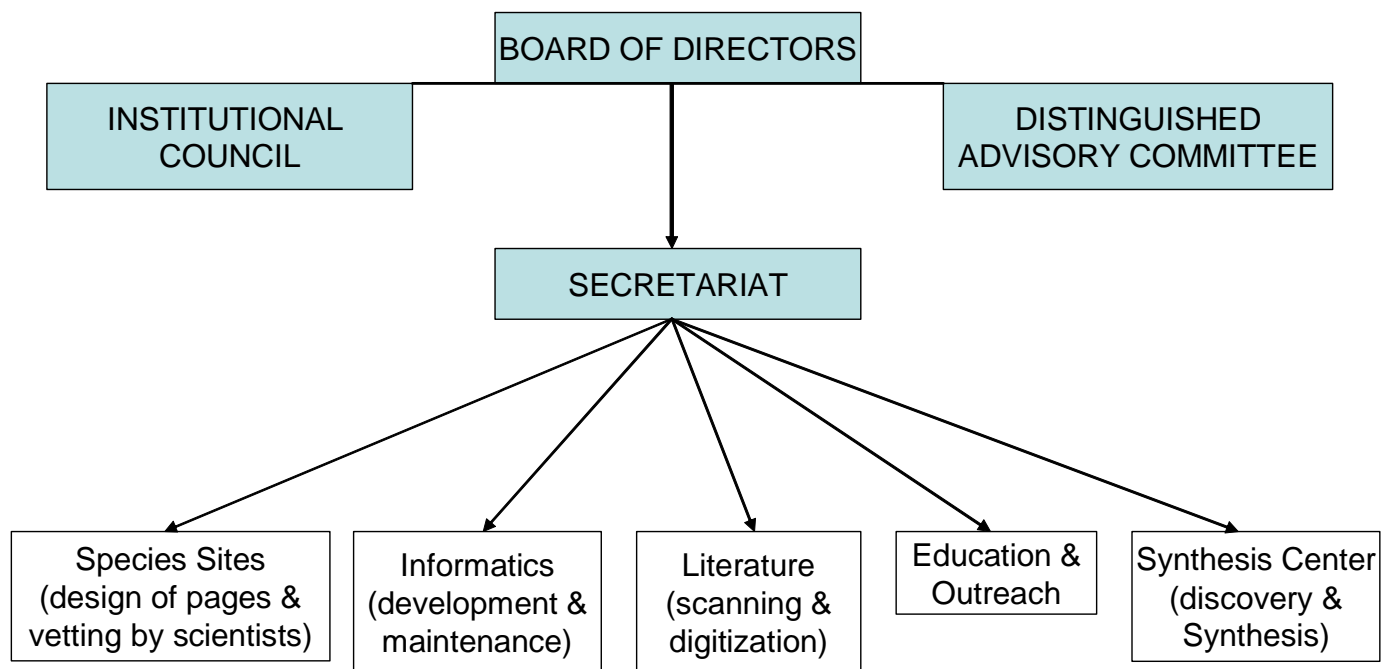
**Planning meetings.** In order to complete the plans for each of the components and broaden the EOL's base of organizations in the US and worldwide, we plan to convene planning meetings during the first half of 2007. These meetings will provide advice in preparing detailed plans and milestones

In addition, each of the five core elements will develop its own stakeholder-based working group(s), to provide feedback from and connection to the relevant communities.

**C. GOVERNANCE AND MANAGEMENT**

The governance and management structure of the EOL is depicted in the diagram below.

**GOVERNANCE AND MANAGEMENT OF THE EOL**



**Board of Directors.** This board will consist of senior figures from the cornerstone institutions (currently the Smithsonian Institution, Marine Biological Laboratory, Field Museum of Natural History, Harvard University, and the Biodiversity Heritage Library Consortium), a legal expert, and liaisons from EOL’s financial sponsors. It will provide hands-on guidance and be the major forum

for coordination of the various facets of the project. The Board is accountable for EOL's success and meeting performance metrics. Over time, other institutions that make major commitments or assume large responsibilities might join the Board of Directors.

**Distinguished Advisory Committee.** This body will consist of 10-12 distinguished individuals, from within and outside the biodiversity community. The Committee will help develop and advocate for the EOL vision, and provide high-level, expert guidance. It will balance political geography, taxa, marine and terrestrial interests, and producers and users of knowledge.

**Institutional Council.** The Council will engage the wider international and scientific communities by providing representation for 15-20 institutions and programs, from the US and other countries, that are crucial partners committed to implementing the EOL. Members will serve on behalf of their institutions or programs. The Council's mandate is to bring a wider perspective for and involvement in the EOL, identify individuals and programs that should be involved in the project, and engage additional funding and other resources.

**Secretariat.** The EOL Secretariat will be hosted at the Smithsonian's National Museum of Natural History. Led by an experienced Executive Director, it is responsible to the Board of Directors for managing the project. In addition to coordinating all aspects of the EOL, it will also be the major site for designing the pages for the species sites, for communicating with and involving the relevant scientific communities, and for overseeing the editorial control of the species pages.

Although all of the cornerstone institutions will be involved in all of the core components of the EOL, they will also act as leaders for specific components, as follows:

Smithsonian Institution: host for the EOL secretariat, and co-leader (with Harvard University) of the Education and Outreach core component

Marine Biological Laboratory (MBL): development and maintenance of aggregation and related software, including the portal for entering the EOL

Biodiversity Heritage Library (BHL) consortium: scanning and digitization of literature

Field Museum of Natural History (FMNH): home to the Biodiversity Synthesis Center and host of many EOL meetings and conferences

Harvard University: co-leader (with the Smithsonian) of the Education and Outreach component

#### **D. INDICATIVE BUDGET FOR THE FIRST FIVE YEARS OF EOL**

The budget for the first five years of operation of the EOL is estimated at \$40 million. Our goal is to secure a lead gift from one or more private foundations for at least half of this budget. The cornerstone institutions have agreed to work together, and with other national and international partners, to raise the remaining funds.

#### **E. PROVISIONAL MILESTONES FOR THE FIRST TWO YEARS OF THE PROJECT**

The milestones in the tables below presume that the project is started in early 2007 and that the funding for the major components of the project is in place by June 2007. Clearly the intensity and speed of the effort will be dependent on success in fundraising.

<b>Milestones for 2007</b>	<b>Lead institution(s)</b>
Establish Board of Directors and begin monthly meetings	Board of Directors
Establish Secretariat; hire key staff	Smithsonian
Convene expert meetings to flesh out plans for the five major EOL components (species pages, software and hardware, digitized literature, education, synthesis center)	Cornerstone institutions
Determine immediate hardware/software needs and initiate purchases of off-the-shelf items	MBL, Secretariat
Hire programmers to begin tailoring aggregation and wiki software	MBL
Identify relevant international partners and invite them to join the project	Board of Directors, Secretariat
Recruit Distinguished Advisory Committee	Board of Directors
Recruit Institutional Council	Board of Directors
Establish first regional center for scanning literature	BHL

Design team begins developing templates for entry-level species pages	Secretariat
Design team establishes working group representing different stakeholder groups	Secretariat
Identify “low hanging fruit”, i.e. taxonomic groups that already have good species pages that can be retrofitted into the EOL format	Secretariat, MBL
Establish first agreements with data providers to incorporate their information into EOL	Secretariat, MBL
Use aggregation technology to prepare first drafts of entry-level species pages	MBL
Wiki tool for experts use in scrutinizing draft pages is in place	MBL
Discussions have taken place with scientific societies, relevant international organizations (e.g. Species 2000) and other groups of experts to identify mechanisms and individuals for reviewing the draft content on the species sites	Secretariat (in concert with FMNH and other partners)
Digitizing of literature begins—by end of year, at least 5 million pages have been scanned	BHL
Taxonomic intelligence tools are used to identify scientific names in the digitized literature and to put these names into the broader taxonomic context	MBL, BHL
Formal and informal education communities begin to examine ways in which the species sites can be used	Harvard, Smithsonian
Synthesis Center is established and key staff hired	FMNH
Synthesis Center has held organizational meetings to identify most promising approaches for using the species sites in synthetic research	FMNH
Hold the first international EOL conference by circa end of 2007	Board of Directors, Secretariat

<b>Milestones for 2008</b>	<b>Lead institution(s)</b>
Design group has completed user needs analysis for entry-level species page	Secretariat
At least 70,000 draft species pages are in place	MBL
First version of EOL workbench has been prepared and is in beta testing	MBL
At least 35,000 species pages have been examined by experts and are ready to serve via the portal	Secretariat (in concert with FMNH and other partners)
Second regional center for digitizing literature is brought on line	BHL
At least 25 million pages of literature have been digitized by the end of year	BHL
Recommendations for incorporating EOL into formal and informal education have been drafted—some model curricula have been developed	Harvard, Smithsonian
First synthetic data analyses have begun, and projects on developing consensus taxonomies of understudied groups are underway	FMNH

## F. OUTCOMES AND EVALUATION

On a scientific level, EOL will accelerate learning answers to essential questions in biodiversity, including startlingly basic ones such as the numbers of known kinds of life and fresh ones in comparative biology that vastly enlarging databases just now allow us to imagine.

On a practical level, the outcomes should be improved management of natural resources, for example with regard to designation of marine protected areas.

On a broader societal level, the great outcome is access to information, which we expect will lead to better informed publics worldwide and will involve those publics in citizen-science activities, such as documenting biodiversity.

**Evaluation.** Many *outputs* of the EOL can fortunately be measured in a timely and quantitative way. The number of species sites, number of literature pages digitized, number of users of the EOL and the number of scientific papers that cite the EOL can be monitored on a regular basis.

*Outcomes* are of course harder to assess. Using proxies such as papers, citations, and prizes, we can assess outcomes within science. Surveys of decision-making or case studies about protection of the environment and use of natural resources could seek to identify the contribution of the EOL. With regard to general public knowledge, it is far easier to assess web traffic than to assess what persons actually know or understand. One might look to changing levels of support for biodiversity protection and research as an indicator of success to which EOL contributes, as well as involvement in citizen science projects.

Overall, we do not think EOL will be difficult to evaluate. If successful, EOL will be famous and heavily used. If after 5 years EOL remains obscure and hard to use, a few interviews will reveal the problems. We would propose an external review by a small group representative of different stakeholders after about 2.5 and 4.5 years.

**Impacts.** At a lofty level, the impacts are improved knowledge contributing to the protection of nature and human well-being. How information is finally used is hard to anticipate and cannot be controlled. Reliable information that can be used for identification can reduce the commercial traffic of endangered species or reduce the number of invasive species that can have devastating impacts on agriculture and human health. Information about species may help in developing field guides for ecotourism projects, an important economic sector that is growing fast worldwide. Providing access to biodiversity can empower researchers in developing countries to have a greater impact on international science, while empowering students and the general public to contribute information to the EOL can bring a wider corpus of data to help us gain new insights into the natural world.

The EOL probably stands to benefit most those at the two ends of the information spectrum. On the one hand, the EOL radically democratizes access to information, so that people in countries with poor libraries and museums or little cash to buy books might suddenly have access to terabytes on their cell phones. On the other end, those with the largest screens and bandwidth, including scientists, will also revel in EOL.

## **G. LONG-TERM SUSTAINABILITY**

The EOL is completely consonant with the basic mission of natural history museums and herbaria. These institutions, collectively, are already a kind of encyclopedia of life, but one that is very difficult to access and search. The EOL is the inevitable digitization and integration of key aspects of these institutions. As such, several of them have accepted that it is their responsibility to lead the creation of the EOL and to find ways to sustain it.

All the partner institutions in this proposal are committed to raising additional funds to realize the vision of the EOL. Over the longer term, they hope to build the costs of maintaining the EOL into their core functions.

The software and knowledge generated by the EOL will be designed for community ownership. The software will be modularized and placed in an Open Source environment so that everyone can add to it and improve it.

No one can anticipate very far ahead the rapidly evolving pattern of revenue (or expense) associated with the Internet. The EOL partners will explore a range of business models. These could include basic subventions from institutional budgets, winning of grants for particular functions or enhancements, revenues from advertising or sponsored links, other kinds of corporate sponsorship, individual memberships and donations—in short the full range of revenue strategies already successfully used.

The EOL will earn social sustainability as the various user bases come to rely on the information it provides. In particular, the educational components (including both formal and informal education and citizen science) and the outputs from the Biodiversity Synthesis Center will contribute to the involvement of the general public and the scientific community and their insistence that EOL be maintained. As GenBank has enabled a wide range of economic activities and is seen as a public good worthy of continuous funding from NIH, there is a real possibility that EOL will achieve such status in the eyes of other agencies, as the data and information to which it will provide access are at the base of a wide range of economic, social and intellectual endeavors.

Appendix 1—E. O. Wilson’s article, “The encyclopedia of life”, from *TREE*, **18**(2)

# 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 expandable. 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 © 2002 Elsevier Science Ltd. All rights reserved. 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<sup>-1</sup>, 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

<http://www.trends.com>

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.

<http://trends.oxfordjournals.org>

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/moritz06moritz.html>
- 13 Sten, L. (2002) Creating a bioinformatics nation. *Nature* 417, 119–120

## **Appendix 2—Rationale for the Encyclopedia of Life**

### **Why is the idea important?**

There are several reasons why we believe it is important to have an EOL.

First, access to information is arguably the defining revolution of our times. As a result, scientific literacy has never been more vital to new generations. The EOL is predicated on the idea that free access to recorded knowledge about life should be encouraged and enabled. It thus exemplifies the Creative Commons (<http://creativecommons.org/>) and its offshoot, the Conservation Commons ([www.conservationcommons.org](http://www.conservationcommons.org)). The information provided by the EOL will be of particular value for students, scientists and people with limited means of access because of income or location, including individuals in developing countries.

Second, popular demand for biological information is enormous, as evidenced by traffic to pages about sharks or sales of field guides of birds. Commercial sources satisfy some of this demand, and the EOL might create opportunities for value-added software by private vendors, just as private companies provide custom services on top of the forecasts of the US National Weather Service. However, the fraction of the 1.8 million species for which commercial field guides historically have been viable is tiny and they are often only available for selected countries and languages.

Third, resource managers all over the world need access to species-level information to make decisions about, for example, trade in endangered species, detection of invasive species, declarations of quarantines, and preparation of environmental impact statements

Fourth, the accumulation of the pages of the EOL will bring an explosion in the field of comparative biology by allowing researchers to examine information encompassing entire

taxonomic groups, such as all scarabs, and to match and layer information across taxa, for example about frogs and flies, or tuna, squid, and zooplankton. It will also link this information with data from biological collections, enabling comparisons of populations and species.

Fifth, the pages will have important uses for the general public and in education. They will be used to enable new kinds of citizen science. For example, they will provide a framework for natural history societies and related citizen groups to undertake local inventories of the flora and fauna in their area. Over time, adding these local efforts together will produce a national inventory of biodiversity which can be used in modeling a wide range of phenomena, from climate change to human impacts. And the pages will also be of great value in formal education, at the middle-school, high-school and college levels, where they can be used to establish biodiversity as a key theme in life sciences curricula.

### **Which other foundations, universities, and agencies are working on the idea?**

The most important complementary large foundation activity is probably Mellon's support to scan and digitize botanical type specimens that were originally collected in Africa and Latin America, thereby facilitating access to the information in the countries of origin. Some small foundations, including JRS and Lounsbery, make modest-sized related grants. The Sloan and Moore foundations provide funding for the Barcode of Life Initiative, involving 100 institutions in 40 countries to develop short DNA identifiers for plants and animals.

The US National Science Foundation funds several large-scale biodiversity efforts, including their programs on Assembling the Tree of Life (determining the phylogenetic relationships of the major branches of life) and Planetary Biodiversity Inventories. But to date it has not been interested in the prospect of amalgamating these data into a single, web-accessible resource.

Biodiversity informatics is an emerging discipline or sub-discipline now growing at several universities and research centers in the US and abroad, for example, U. of California at Berkeley, Kansas, Rutgers, Arizona State University, CONABIO (Mexico), INBio (Costa Rica), the U. of Auckland (New Zealand) and the Atlas of Living Australia (spearheaded by the Australian Biological Resources Study). The Missouri Botanical Garden is a leader in plants, together with other institutions in the iPlants Consortium. The planning group for the EOL has scanned related initiatives and compiled extensive lists of relevant initiatives and interested parties.

Several large compendia of species pages have already been developed, such as FishBase, CephBase ([www.cephbase.utmb.edu](http://www.cephbase.utmb.edu)), AmphibiaWeb (<http://amphibiaweb.org/>), North American Mammals ([www.mnh.si.edu/mna](http://www.mnh.si.edu/mna)), and a host of others. EOL will develop synergistic collaborations with as many of these sites as possible. In addition, the seminal iSpecies site (<http://darwin.zoology.gla.ac.uk/~rpage/ispecies>) uses aggregation (mash-up) technology to generate species pages “on the fly”, much as EOL will do in order to develop the first-cut pages for expert evaluation.

### **How is EOL different from Wikipedia and Google?**

We are full of admiration for both Wikipedia and Google, and their success inspires us to believe we can create the EOL, drawing on key attributes and strategies of each.

The EOL differs from Wikipedia and Google in that it will systematically provide access to expert knowledge. The vast majority of the data available through the EOL will have received scientific scrutiny. In contrast to Wikipedia entries, EOL’s will have standard formats, required fields, and preference for vetted information. In contrast to Google, users will not have to sort out inappropriate material that comes from a Google search. A birdwatcher interested in “Baltimore

Orioles” will find that a Google search results in about 2.5 million hits. Using the scientific name for this species (*Icterus galbula*) still results in 105,000 hits, of varying degrees of authoritativeness.

In addition, Google cannot access deep content on the web. Over 90% of web-accessible content is within databases that Google cannot penetrate. By working with the holders of relevant biological databases, EOL will make their contents available through its species sites.

Another difference from Google is that the EOL is founded on the ideas of open content, open access, and complete transparency. Google’s cached data belong to Google, not the general public. Only Google’s search engine can search Google, and it has lots of biases and advertising. And Google has no long-term commitment to biodiversity as a subject matter. Like Wikipedia, EOL has an open-source, open-content model. Only Google employees make Google products, whereas EOL will have publicly accessible boards of editors, a wiki-type workbench that allows experts to add information directly, and transparent operation.

### **How will the EOL deal with Intellectual Property Issues?**

The advocates of the EOL are keenly aware of the many and dynamic intellectual property issues associated with the venture and have examined these during the past year. The EOL would expect to be a fast follower, not a pathbreaker in regard to Internet practice. We expect most law actions and norms will follow from practices of Google, YouTube, mySpace and other entities, private and public, that are pushing the envelope of practice.

As a basic position, EOL would expect to quickly take off-line material that becomes legally controversial until practices are sorted out. EOL will include on its Board of Directors an individual with expertise in relevant law. For example, the UC Berkeley law clinic includes several faculty with keen interest in this area who might serve and assist at no or minimal cost.

**How long will it take to make substantial progress?**

Within two years the EOL website should be largely enabled and synthesizing large numbers of provisional pages. Within three years several exemplary groups should be dynamically available at a high level of quality. Within five years, as many as 1 million species should be accessible and substantial progress will have been made in using this resource to synthesize new knowledge. The remaining 800,000 species sites relate to lesser-known groups and may take five more years work.

### **Appendix 3—EOL Technical Vision**

The EOL project is grand in scale and scope and must be a communal enterprise, leveraging and integrating the efforts of other initiatives to engage authoritative data sources and experts. The informatics component of EOL will contribute to the challenge of creating a web site for every species by building a new kind of cyberinfrastructure that can draw together the growing body of information that is distributed across thousands of web-sites. The cyberinfrastructure has three innovative components, all of which have been explored using prototypes developed by the MBL and other partners. The first is ‘Taxonomic Intelligence’ – an approach that addresses unique problems associated with information about organisms. The second is the use of aggregation technology (mashups) to ‘atomise’ data from collaborating web sites and to recompile the information into new web pages. The third component is a layered architecture that is flexible, powerful, and promotes a modular approach to software development, so fostering communal ownership. Although some components of this cyberinfrastructure have been developed, a number of elements have yet to be assembled, including concept management, a comprehensive indexing and navigation system, and a modular environment through which experts can work with distributed content.

#### **TAXONOMIC INTELLIGENCE**

From the time of Linnaeus, biologists have used a system of names for organisms that seeks to ensure that each species has a unique but standardized name. Virtually all biological data systems from books to on-line databases rely on names. Names offer the principal means by which the EOL can gather together information from many websites on any one species. For various reasons, the

‘each species has a unique name’ goal has not been achieved. There may be different names for the same organism (synonyms), and the same name may be used for different organisms (homonyms).

The components that are being developed to collectively resolve the problems of managing information about organisms are referred to as ‘Taxonomic Intelligence’ (TI), which will continue to evolve as needs and technologies allow for increasing sophistication. Currently we think of TI as including

- A comprehensive list of all species names and names surrogates of all organisms in all languages as the foundation of indexing
- Dealing with the synonym problem through reconciliation
- Dealing with homonyms and concepts through disambiguation devices
- An inclusive use of taxonomic hierarchies to enhance searches and browse content within a biologically meaningful structure

## **AGGREGATION TECHNOLOGY – KEY TO THE WORKBENCH AND THE PORTAL**

The informatics aspect of EOL has to collate elements of information from an unlimited array of web-sites. This requires selection of existing pieces of information and their association and recombination in different ways to meet the needs of a diverse audience. To do this, we propose to employ Aggregation Technology (also referred to as web aggregation, mashup, or web application hybrids). Aggregation Technology uses an understanding of the architecture of source databases or output from those databases to segregate different categories of information. Complementary elements of different web sites can then be grouped together, much as generic search engines can show all images separated from the context in which they were first presented. In this form, the information can be made available in the workbench or portal, where experts and users can select their preferred content, or annotate it with their own metadata. The selected elements can then be

recombined into new pages. Full accreditation of creators and sources are critical to this approach, as is a system for peer review. It is important that in applying this technology, we place minimum burden on the data-providers. This obligates us to find site-specific solutions appropriate to each data provider.

## THE LAYERED ARCHITECTURE

EOL requires an organizational architecture that can assemble facts about any aspect of the biology of any organism into critically reviewed pages that will meet the needs of a diverse

community. Given the success of

‘layered architecture’ to underpin

the operations of the Internet, we

propose to use a similar approach

in which successive layers

convert factual information (the

lowest layer) into the

authoritative and sophisticated

statements required by EOL.

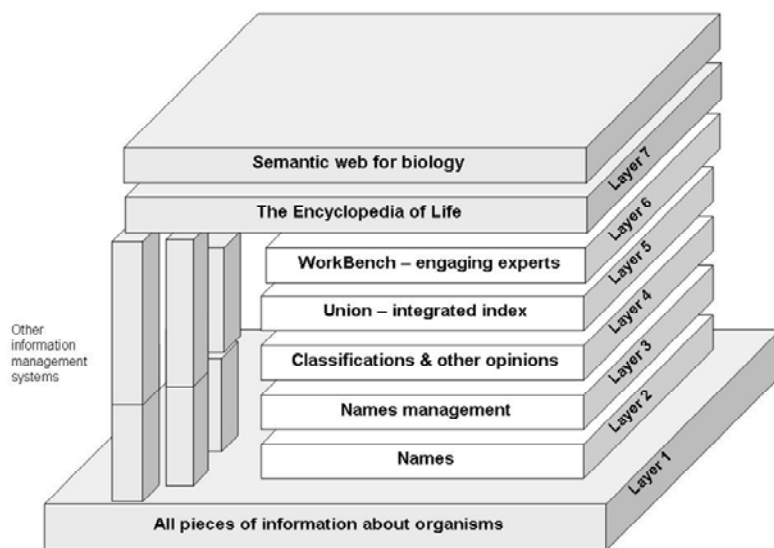
Each layer is defined by unique

properties that determine how it adds value to information. By separating functions and defining the communications interfaces, EOL’s stratified architecture will help to break the informatics

challenges into parts, and is intended to foster a modular approach to software development that will

allow evolution without requiring modifications throughout the entire structure. Modular software

can be communally developed and owned and so is designed with an ever-evolving character. This



Layered architecture

design fosters partnerships with a wide spectrum of data and service providers allowing them to plug into any layer of the architecture.

We conceive of the basal layer as all records about organisms that can be accessed through the internet – that is, the totality of facts upon which we depend. These facts often occur in environments where they are placed in a context, and for the purpose of EOL, the factual content will need to be identified and segregated from context or opinion. Layer 2 includes all of the names and name surrogates that were associated with the factual content. These names form the foundations of the indexing system for EOL, and will be built from the uBio NameBank, currently the largest compendium of names, and other sources. As names are imperfect metadata, the following layer (layer 3) adds the sophistications that allow data on the same species but labeled with different names to be brought together as well as segregation of data attached to the same name but referring to different species. These tasks, addressed through reconciliation and disambiguation, are mostly embedded within the uBio NameBank environment.

The fourth layer is the component that manages the various taxonomies and other classification schemes that exist. As these are all incomplete, the subsequent layer, Union, is an integrated index that accommodates all names and allows content to be viewed in the context of any classification.

The WorkBench (layer 6) is conceived as modular software environment that allows users to see all of the content in the lower layers. Through the WorkBench, users will be able to annotate names, add new names, create their own classifications, and leave comments for the custodians of the name information for various groups of organisms. They will also be able to use WorkBench to create their own desktops or web pages. Our vision is that this layer will be synchronized with major aggregators, nomenclators and other names providers. By having a universal and flexible indexing

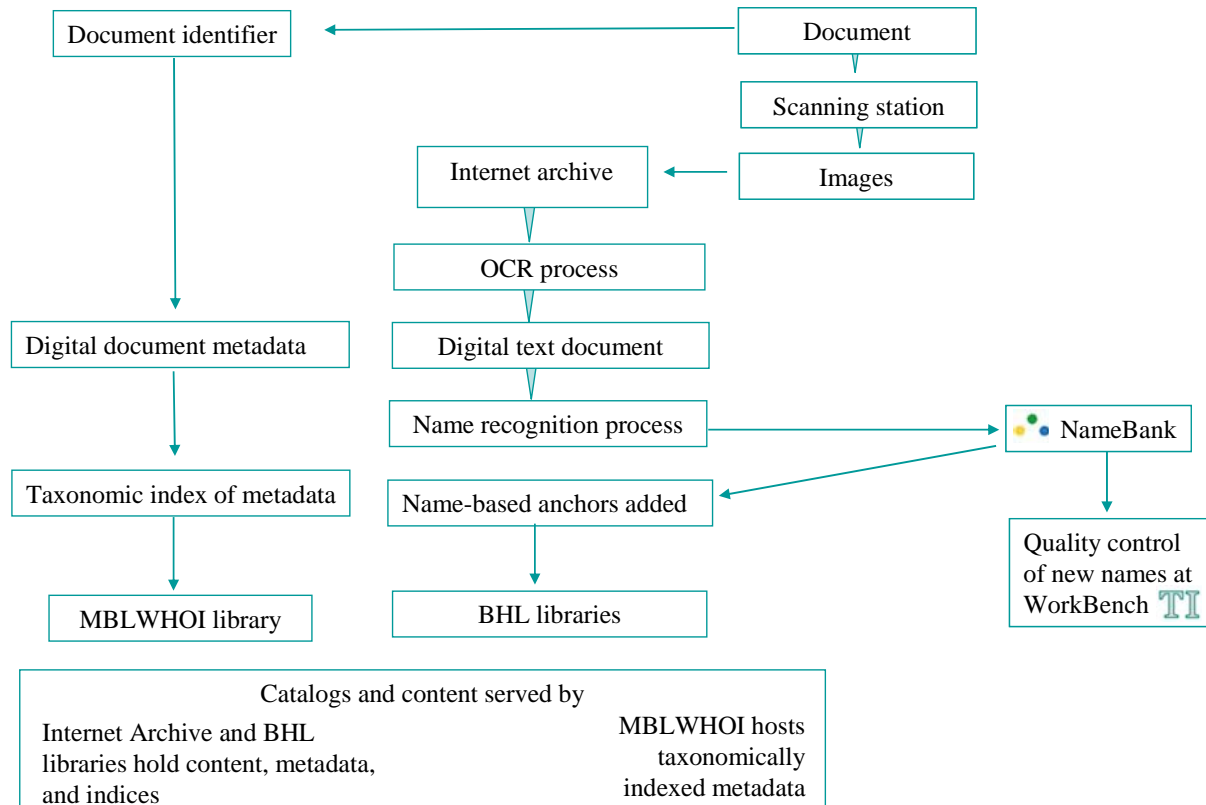
system at its core, WorkBench will thus be capable of serving a great wealth of needs. It will be available universally, and not only to individuals associated with EOL. The modules will be freely available through a Source Forge type environment, in order to engage the community and to imbue the WorkBench with an ever-evolving character. Because of the scope of potential uses, we envisage this as being the most challenging and rich area of coding. For the WorkBench to be successful, it must engage a high proportion of the community of experts.

## **THE PORTAL**

The Encyclopedia of Life Portal (layer 7) is the name given to the environment through which information on organisms will be delivered to end users. The EOL Portal will be developed as a collaborative venture involving the EOL Secretariat, which will be responsible for design; and the MBL, which will be responsible for the 'back end', and a myriad of other partners. The back end will be developed in the same way as the WorkBench, largely as a system of modules that can be selected by the end user to manipulate the content in a form that suits their needs. As with the WorkBench, the software will be placed in an open source environment. Again, aggregation technology will be in play as the means of drawing together information from multiple sources, but now enriched with annotations from the WorkBench. This environment will be compliant with data and communication standards emerging from the Taxonomic Database Working Group (TDWG); and will apply Globally Unique Identifiers (GUIDs) to facilitate re-use of the content by others.

## **INDEXING CONTENT FROM THE BIODIVERSITY HERITAGE LIBRARY**

The EOL informatics team will work in close liaison with the multi-institutional Biodiversity Heritage Library and the Internet Archive to provide taxonomic indexing software for the literature that the BHL will scan and digitize. Enhanced Name Recognition algorithms will be built by combining TaxonGrab, developed at the American Museum of Natural History, with the FindIT



algorithm developed at the MBL. The process will generate a taxonomic index of BHL content. The index will be served from the MBL and mirror sites, the content from the Internet Archive and from the BHL libraries that provided the original material. The EOL Informatics team will continue to develop tools that integrate the content of the BHL within other services.