Science | DOI:10.1145/1562764.1562771

David Lindley

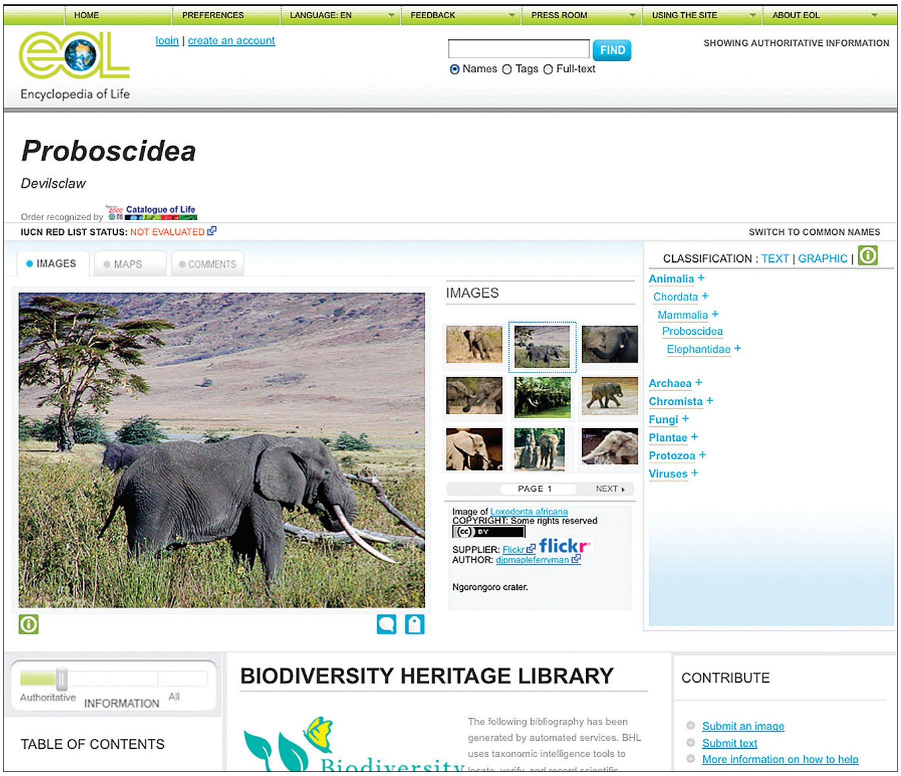
Managing Data

Dealing with terabytes of data is not the monumental task it once was. The difficult part is presenting enormous amounts of information in ways that are most useful to a wide variety of users.

The Herbarium at the University of Alaska's Museum of the North houses one of the world's largest collections of arctic plant specimens and Russian flora. For scientists studying the ecology of this region or the changing biodiversity caused by human encroachment and climate change, it's an invaluable resource. However, the Herbarium is not ideally located for most scientists. Because of the considerable expense of traveling to Alaska, scientists often have specimens temporarily shipped to their institutions, but that also costs money, and delicate specimens suffer from the attendant wear and tear. As a result, the Herbarium is scanning and digitizing its extensive collection, making the images and text available on the Internet to scientists, not to mention enthusiastic amateurs, everywhere in the world.

The amount of data involved—about five terabytes so far—is hardly intimidating by today's standards, but it proved to be overwhelming for the Herbarium's previous database partner. Consequently, the Herbarium teamed up with the University of Texas at Austin's Texas Advanced Computing Center (TACC), gaining access to TACC's Corral, an online 1.2 petabyte data repository, via the National Science Foundation's TeraGrid cyberinfrastructure. "We take images and they're immediately downloaded to Texas, and in live time we have a link to that file from our database," says Steffi Ickert-Bond, curator of the Herbarium and an assistant professor of botany at the University of Alaska. The Herbarium's holdings are now accessible via Arctos, an online collaboration of five U.S. university museums.

Nearly 80,000 of the Herbarium's 220,000 specimens have been digitized by hand, with students laboriously keying in detailed information from the specimens' labels, some of which are handwritten and some of which are in Russian. To record the remaining 140,000 specimens, however, the proj-



Proboscidea

Devilsclaw

Order recognized by Catalogue of Life

IUCN RED LIST STATUS: NOT EVALUATED

Image of Loxodonta africana
COPYRIGHT: Some rights reserved
(cc) BY
SUPPLIER: Flickr
AUTHOR: djmapleferryman

Ngorongoro crater.

The following bibliography has been generated by automated services. BHL uses taxonomic intelligence tools to locate, verify, and record scientific

The ambitious Encyclopedia of Life aims to create a Web page for every known species on Earth.

ect is shifting to automated label scanning using Google's Tesseract optical character recognition engine.

Handling varied data, such as text, numbers, images, and sounds, no longer poses any great technical difficulties, says Chris Jordan, who is in charge of data infrastructure at TACC. Corral uses Sun Microsystems' Lustre file-handling program to present users with a seamless interface to the data. What's trickier is building a presentation layer that gives scientists in a particular discipline access to the resources in an intuitive and user-friendly way. Unlike researchers in physics and engineering, for example, those working in museums or the humanities aren't accustomed to using computers at the command line level, says Jordan, so "a lot of our effort now is on [building] interfaces for people to locate data along with descriptive metadata in a way that's reasonably easy."

Accessible and Useful

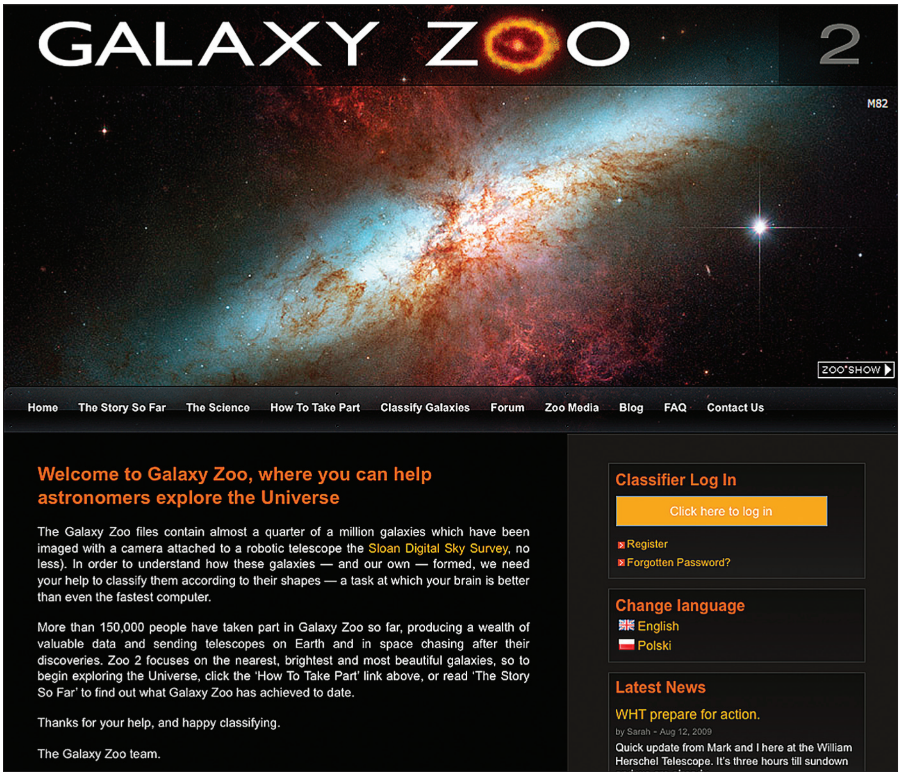
Making vast repositories of biological information widely accessible and useful is the aim of the Encyclopedia of Life (EOL), an ambitious project whose long-term goal is to create a Web page for every known species on Earth. "We want to provide all information about all organisms to all audiences," says Cyndy Parr, EOL's species pages group leader, who is based at the Smithsonian's National Museum of Natural History (NMNH) in Washington, D.C. So far, EOL has about 1.4 million species pages, but most of them are little more than placeholders for known organisms, frequently directing visitors to an external link with more information. Some 158,000 pages, however, contain at least one data object that's been scientifically vetted by EOL. The project is chasing a moving target, since the number of species on the planet is unknown. A figure of 1.8 million known species is commonly accepted, says Parr, but new species are being discovered at a rate of 20,000 or more each year, and some extrapolations suggest that there may be as many as 10 million distinct species.

Far from competing with efforts such as those by the University of Alaska's Herbarium, EOL aims to build species pages that deliver essential information in a uniform style and lead visitors who want to dig deeper to more specialized, detailed resources. Indeed, says EOL Executive Director Jim Edwards, also at NMNH, a principal motive behind EOL was the realization that many research communities are building their own databases—such as AntWeb, FishBase, and others—each with its own design, interface, and search procedures, created to meet the needs of its particular community.

Launched with grants from the MacArthur and Sloan foundations, EOL receives support from several museums and other institutions. EOL invites scientific contributions from amateurs and academics, but uses a network of researchers to decide what information will be included. The site tries to steer a middle course between a pure top-down model, with pages created only by experts, and a self-policed wiki.

EOL resides on servers at the Smithsonian and the Marine Biological Laboratory in Woods Hole, MA, but since it is mainly an index to other information, the data cached on those servers amounts to a few hundred megabytes. EOL's informatics challenges derive in large part from the historical origins of biological information. Even the formal names of species can be treacherous, since some species have been "discovered" on more than one occasion and received duplicate names, and sometimes molecular or DNA analysis demonstrates that what's long been regarded as a single species is, in fact, two distinct species.

In addition, it's essential to be able to search in less formal ways, such as by habitat type, mating behavior, or leaf shape, characteristics that aren't often described by a standardized set of terms. That's a particular issue for one of EOL's partner efforts, the Biodiversity Heritage Library (BHL), a collaboration of 10 natural history museum libraries, botanical libraries, and research institutions in the U.S. and the U.K. that has put nearly 15 million digitized pages from 37,000 books online. Although optical character recognition software has become reliable, scanning millions of pages is labor-intensive and time consuming, says Chris Freeland, BHL's technical director and manager of the bioinformatics department at the Missouri Botanical Garden in St. Louis. Someone—usually a volunteer student—must turn the pages of every book and correctly position them for the camera. After that phase, though, the digitizing process is automated and efficient. Typewritten or commercially printed material is optically well



Welcome to Galaxy Zoo, where you can help astronomers explore the Universe

The Galaxy Zoo files contain almost a quarter of a million galaxies which have been imaged with a camera attached to a robotic telescope the Sloan Digital Sky Survey, no less). In order to understand how these galaxies — and our own — formed, we need your help to classify them according to their shapes — a task at which your brain is better than even the fastest computer.

More than 150,000 people have taken part in Galaxy Zoo so far, producing a wealth of valuable data and sending telescopes on Earth and in space chasing after their discoveries. Zoo 2 focuses on the nearest, brightest and most beautiful galaxies, so to begin exploring the Universe, click the 'How To Take Part' link above, or read 'The Story So Far' to find out what Galaxy Zoo has achieved to date.

Thanks for your help, and happy classifying.

The Galaxy Zoo team.

WHT prepare for action.

by Sarah - Aug 12, 2009

Quick update from Mark and I here at the William Herschel Telescope. It's three hours till sundown

Galaxy Zoo uses crowdsourcing to help categorize galaxies as either spiral or elliptical.

recognized, although unusual scripts, such as older typefaces and cursive characters, can be problematic.

To date, BHL's digitized collection totals 50 terabytes. That's not so large, Freeland notes, but replicating it at mirror Web sites and ensuring its reliability and security is a challenge. Once again, though, the larger headache is presenting the data. Searching text for keywords and phrases is easy, but because a great deal of biological information is qualitative and descriptive, it's very difficult to construct a search that will dig out all references to, say, the mating behavior of bearded dragon lizards. As a result, EOL is working with computer scientists on natural language processing software that can intelligently characterize the meaning of digitized texts. But accomplishing that goal automatically and reliably remains elusive, Edwards says.

Crowdsourcing Galaxies

EOL is also investigating crowdsourcing methods in which a Web site visitor is asked to supply keywords for an image or text extract. One example of crowdsourcing paying scientific dividends is the Galaxy Zoo, an offshoot of the Sloan Digital Sky Survey, which uses an automated telescope to scan the sky for galaxies and digitize images of them. (For more about the Sloan Digital Sky Survey, see "Jim Gray, Astronomer" in the November 2008 issue of *Communications*.) The Sloan survey examines so much space that astronomers cannot hope to look at every galaxy image, yet direct visual inspection is how astronomers have traditionally gained their understanding of galaxies, says Bob Nichol, a professor of astrophysics at the University of Portsmouth in England. Based at the University of Oxford, the Galaxy Zoo gets some of that irreplaceable scientific insight from Web site visitors, who are asked to classify a series of galaxy images drawn randomly from the Sloan survey as spiral or elliptical. The task can be addictive, says Nichol, and 250,000 visitors have collectively classified nearly one million galaxies at least 30 times each. One statistically solid result to emerge is that 15% of galaxies don't obey the usual rule that the color of elliptical galaxies tends toward the red end of the spectrum and spiral galaxies are bluer. Moreover, red spiral galaxies tend to inhabit the outskirts of galaxy clusters, a finding that allows cosmologists to test theories about the galaxies' origins.

The Encyclopedia of Life is working with computer scientists on natural language processing software that can intelligently characterize digitized texts.

One crowdsourcing lesson Nichol has drawn from the success of the Galaxy Zoo is that it is crucial to precisely tell visitors what information is needed and why it is important. Some Galaxy Zoo scientists have been discussing with other researchers, including botanists, how to translate their methods to other disciplines, but for a project as large as EOL, the task is daunting. The required information is not easily reduced to a handful of simple questions, and while pages for popular mammals and pretty flowers receive many hits, pages for undistinguished flies and obscure bacteria languish unseen in the cyberspace equivalent of a dusty museum attic. Crowdsourcing doesn't work without crowds.

Although EOL has to date attained only a tiny fraction of its ultimate goals, it has earned respect and enthusiasm from those in the biological community who were initially skeptical about its utility, Edwards says. And as EOL continues to grow, he thinks its value will be understood by many different communities, including the public and commercial businesses, so that it will become an indispensable resource. C

David Lindley is a science writer and author based in Alexandria, VA.

Calendar of Events

October 15–16
2009 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement,
Lake Buena Vista, FL
Sponsored: SIGSOFT
Contact: Laurie Ann Williams
Email: williams@csc.ncsu.edu

October 17–18
First Asia-Pacific Symposium on Internetware,
Beijing, China
Contact: Hong Mei
Email: meih@pku.edu.cn

October 18–20
4th International Conference on Nano-Networks,
Lucerne, Switzerland
Contact: Alexandre Schmid
Email: Alexandre.schmid@epfl.ch

October 19–20
Symposium on Architecture for Networking and Communication Studies,
Princeton, NJ
Contact: Peter Z. Onufryk
Email: peter.onufryk@idt.com

October 19–24
ACM Multimedia Conference,
Beijing, China
Sponsored: SIGMM
Contact: Wen Gao
Email: wgao@pku.edu.cn

October 20–22
4th International Conference on Performance Evaluation Methodologies and Tools,
Pisa, Italy
Contact: Giovanni Stea
Email: g.stea@iet.unipi.it

October 21–23
Asia Information Retrieval Symposium,
Sapporo, Japan
Contact: Tetsuya Sakai
Email: tetsuyasakai@xa2.so-net.ne.jp

October 22–24
ACM Special Interest Group for Information Technology Education Conference,
Fairfax, VA
Contact: Donald Gantz
Email: dgantz!@gmu.edu