Fossil Databases Move to the Web

Paleobiologists readily acknowledge that they lag behind disciplines such as molecular biology in sharing data on the Web. But several researchers are working to put existing data sets online. And at least one team hopes to build a sort of GenBank of paleobiology, a Web site where everyone can deposit their fossil finds. Organizers of these efforts face a big hurdle, however: deep divisions within the community over the sharing and quality of data.

Although a Web search will turn up at least a dozen paleo databases, many consist merely of photos of selected fossil specimens or taxonomic lists. A few broader databases limited to one region, animal or plant group, or time period can be downloaded from the Internet in one huge chunk. What's missing are comprehensive, open-access, interactive Web databases that archive published data on where and when a species lived, information that is critical for analyzing patterns of evolution and extinction.

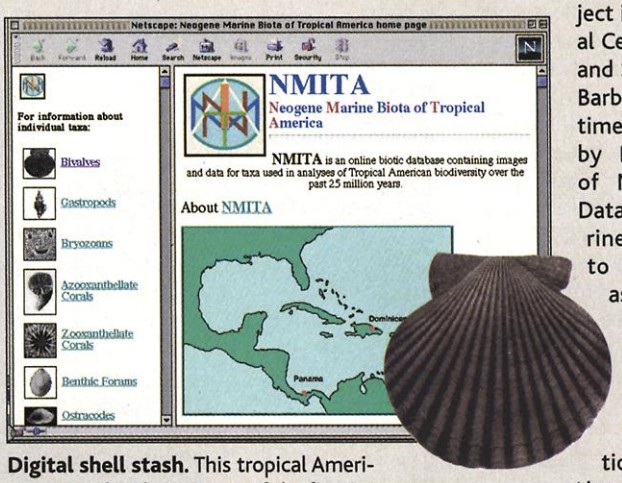
Paleobiologists point out that they face the difficult task of integrating species, temporal, spatial, and geochemical data that can quickly become obsolete if new fossils are discovered. "It's a much more complicated endeavor" than even living species databases, says Doug Erwin, paleobiology curator at the Smithsonian's National Museum of Natural History (NMNH) (see main text).

But would-be Web database builders must also deal with an ambivalence over such repositories that goes back to the most famous one, a marine species database built in the 1970s that the late Jack Sepkoski of the University of Chicago used to overturn many ideas about extinctions and diversification. Besides logging his own fossil finds, "he grazed and browsed in the literature and used it in ways that made some paleontologists uncomfortable," says Kay Behrensmeyer of NMNH. Such attitudes are still common among vertebrate paleontologists, whose fossils are relatively rare. These views play out both as reluctance by some collectors to share their specimen databases and as long-running disputes over the quality of compilations such as Sepkoski's—whether he used correct taxonomy, for example.

Despite that baggage, a few broad Web paleodatabases are under construction. They've been spurred by advances in "relational database" software that make it possible to dovetail separate data sets so that, say, shifting the Eocene period by 1 million years doesn't mean having to adjust every entry in a database. Such tools "made a huge difference," says Charles Marshall of Harvard University.

One impressive experiment in cooperative database building is Neogene Marine Biota of Tropical America, which logs marine fossils from the last 25 million years (porites.geology.uiowa.edu/index.htm). At this stage, the site, which emphasizes photos, is basically "like a Peterson's field guide" for identifying specimens, says co-curator Ann Budd of the University of Iowa in Iowa City. But her group plans to add data that's now accessible only to contributors so that other users can plot ranges and evolutionary trees.

Another project just coming online is the Evolution of Terrestrial Ecosystems (ETE) database developed by NMNH and John Damuth of the University of California, Santa Barbara. It covers both animal and plant terrestrial fossils and includes age, species lists, body size, and diet for nearly 4000 localities, largely from the African late Cenozoic. This week, ETE debuted a pilot Web version of the database (etedata.si.edu). Behrensmeyer hopes more researchers will contribute. "Once people see what can be done, I really think they will be willing to provide access" to their data, she says.



Digital shell stash. This tropical America marine fossil site is one of the first interactive paleobiology Web databases.

An even more ambitious project is under way at the National Center for Ecological Analysis and Synthesis (NCEAS) in Santa Barbara: The goal is to span all time periods and organisms. Led by Marshall and John Alroy of NCEAS, the Paleobiology Database is starting with marine paleofauna but plans to fold in other data such as Alroy's own North American mammalian databases (www.nceas.ucsb.edu/public/pmpd). "We need to have integrated databases to answer big-picture questions" about evolution, says Alroy. An open database where anyone can enter data via the Web "represents the only rational solution," he says. Alroy has found 36 collaborators so far but predicts that some private databases "will never be online."

Indeed, skeptics of the all-in-one database idea abound. Richard Stucky of the Denver Museum of Nature and Science, who's compiling Cenozoic North American mammal data to expand an older database called FAUNMAP, asserts that he's painstakingly gathered data using "strict criteria" to address specific research problems. Asks Stucky: "Can a central database answer all the questions [researchers are] asking? I say, 'No.' "

But others look forward to a day when anyone can troll a central fossil database. Says Marshall, "It seems daft to go into the field and collect a bucketful of fossils and not see it recorded anywhere."

—JOCELYN KAISER

a very fierce streak of independence," notes Kansas informatics expert Jim Beach. Data-sharing is often carefully worked out among researchers studying a particular group of organisms. And often individuals want to wait until they've published all their results before they open their books to colleagues. Complicating matters further, each museum has had its own way of doing things: Some, like Kansas, compiled information in ledgers, others used computer text files, and still others developed spreadsheets.

To Vieglais fell the task of coming up with a system that could work with all these types of collections. He first tested his software in 1998 by incorporating collections from the Kansas herbarium along with the museum's bird and mammal data into the prototype Species Analyst. It was able to retrieve information, despite differences in the formats of the collections. Soon afterward, he added mammals on file at the University of California, Berkeley, and he's been signing up museums, collection by collection, ever since. He lets the curators in charge of these collections decide how much to put online, so they are comfortable with sharing their hard-earned data.

As an added feature, Species Analyst forwards the geographic information about a given species in each collection to the San Diego Supercomputer Center. There, a program called GARP developed by David Stockwell maps that information and, based on the environmental data available for those sites, predicts the species' environmental