# MUSEUM DRAWERS GO DIGITAL

## New technology speeds efforts to display billions of natural history specimens online

*By* **Nala Rogers***, in Washington, D.C.*

In a back room of the Smithsonian Institution's National Museum of Natural History (NMNH) here, Rochelle Safo handles preserved plant specimens with reverence. The brittle leaves and stems, dried and glued to pieces of paper, hold a wealth of knowledge that spans continents and centuries. Every 4 to 6 seconds, Safo places a new one on the conveyer belt so it can be photographed and digitally shared with the world. "On a good day we can do 3500 specimens," says Safo, a digitization specialist who has been working on the project since it launched in October 2015.

The NMNH conveyer belt system is part of a global effort to open up access to museum collections. No one knows exactly how many natural history specimens exist in museums and other research institutions worldwide, but some calculate it's on the order of 3 billion. In most cases, the displays seen by visitors make up a tiny slice of this treasure; museum curators estimate that more than 99% are stored away from the public gaze.

Researchers have for decades used museum specimens to answer questions about how species diverge, where they move around the globe, and how they respond to changing conditions. "There is more information about biodiversity in natural history collections than in all the other sources of information put together, outside of nature itself," says Larry Page of the Florida Museum of Natural History in Gainesville. "But it's been mostly inaccessible." Researchers wanting to study the specimens have traditionally had to travel from museum to museum in person, or else request that the specimens be mailed to them.

Now, even as they struggle with funding woes that limit their activities, institutions from China to Europe to the United States are working to put specimen photographs and related information online where anyone can view them. Until recently, these efforts were slow and painstaking, barely chipping away at the staggering amount of data in collections. Now, technological advances and innovative workflows are allowing institutions to think bigger, ushering in a new age of mass digitization. With their previous system, the NMNH herbarium staff could digitize 30,000 specimens in a year—photographing them and transcribing label data—at a cost of $5 to $7 per specimen, says Sylvia Orli, a botanist and digitization project leader at NMNH. Now, they expect to finish 650,000 in that time, each one costing just $1.

As digitization grows faster and cheaper, more governments and institutions are investing in it. The NMNH conveyer belt is funded by the Smithsonian's 5-year old Digitization Program Office (DPO) here. And since 2011, the U.S. National Science Foundation (NSF) has devoted $10 million per year to digitization efforts in nonfederal collections across the United States—Page, for example, is project director for iDigBio, an NSF-funded effort to coordinate biological specimen digitization. But even with these new funding opportunities, museum officials and curators stress that there is still far too little money to make all such specimens digital.

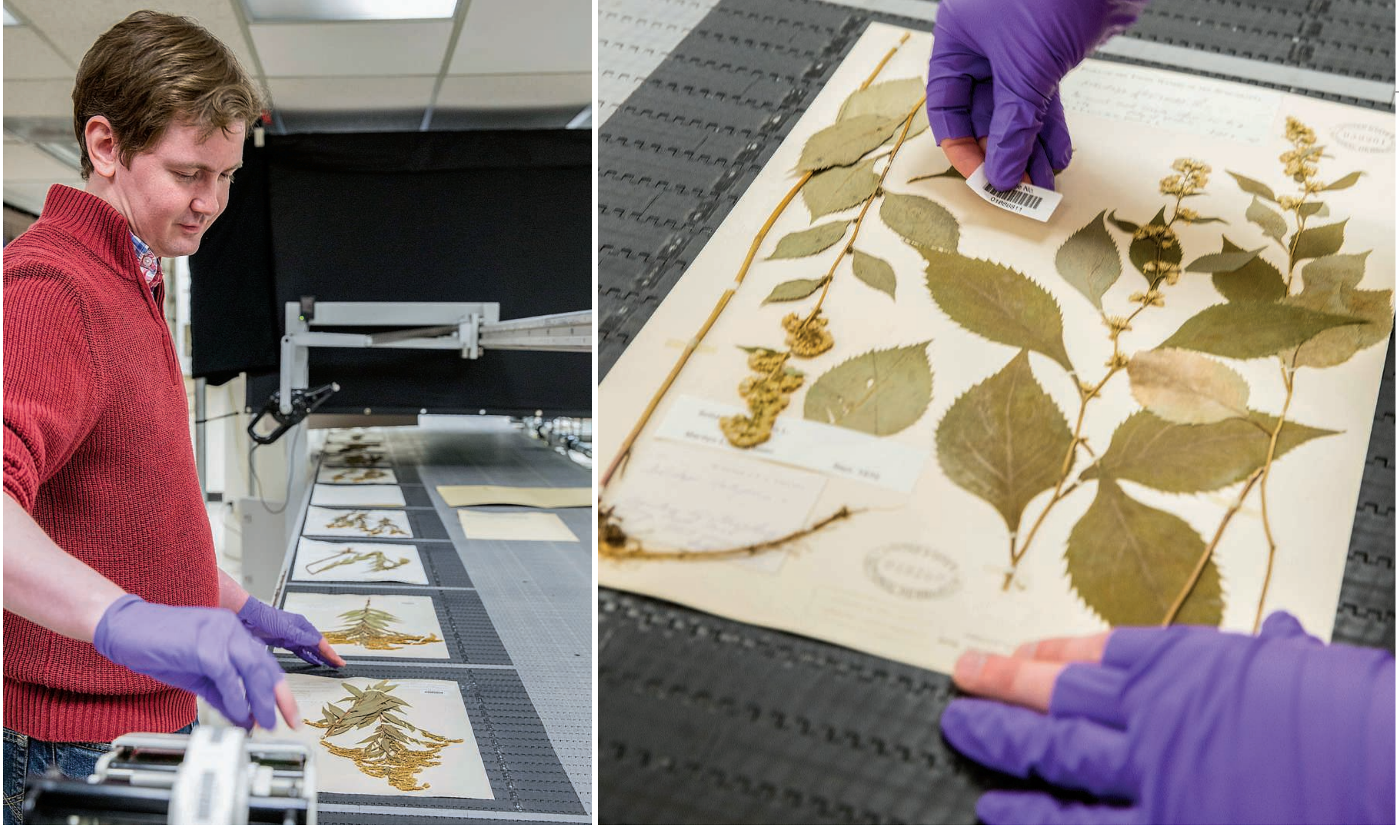
Some recent digitization projects have already born scientific fruit, yielding insights into everything from invasive species to climate change. "I can get into a database and bring up one image after another of a

Staff at the Smithsonian Institution's National Museum of Natural History manage about 5 million specimens of algae and plants, some dating back centuries.

PHOTO: C. CLARK/SMITHSONIAN

Published by AAAS

Published by AAAS



Justin Donaldson places barcode stickers on plant specimen sheets as they travel down a conveyer belt toward a camera at the Smithsonian Institution's National Museum of Natural History in Washington, D.C. The barcodes help software keep track of the images and link them to database records.

PHOTOS: P. FETTERS

plant that I'm interested in," says Michael Donoghue, a phylogenetic biologist at Yale University who co-leads a large digitization project on New England vascular plants. "I can measure the leaves, or measure something about the flowers, and quickly do a scientific study that I never could have done before, because it would have taken me my entire life to go around to every freaking museum."

**EFFORTS TO DIGITIZE** natural history collections aren't new. Curators at NMNH started entering specimen data into computers in the 1960s, when the cutting-edge technology was punch cards, Orli says. The NMNH herbarium began posting records of its most important specimens—the ones used to define species—online in 1992.

Plants and algae lend themselves to digitization because they are usually pressed flat and attached to pieces of paper. A single photograph can capture the label and most of the structural details needed for research. The Paris National Museum of Natural History was one of the first to adopt mass digitization technology with its own conveyor belt system, and it finished off its entire vascular plant collection—about 6 million specimens—in 2012. Other herbaria across the world are working rapidly through their collections with a variety of semiautomated procedures, funded by government grants or the museums themselves.

At NMNH, the new conveyer belt setup has sped up the imaging process by a factor of 10. One person lays herbarium sheets on the 9-meter belt, another attaches barcode stickers so that software can keep track of the images, and a third replaces the sheets in folders when they have finished their journey. Every few seconds, a rumbling chug brings a new specimen beneath the camera.

The images captured are so detailed that researchers can count fern spores less than 50 microns across, says Ken Rahaim of DPO. The plan is to digitize 500,000 specimens, Orli says. After that, unless the Smithsonian can find more funding, the herbarium will have to shut the conveyer belt down and use older techniques to slowly digitize the rest. "It's only a dollar a specimen. You think 'god, that's so cheap.' But we've got 5 million specimens," says Vicki Funk, a botanist at NMNH who uses digitized specimens to study plant systematics.

**2–9 billion**

Total number of natural history specimens, according to several estimates.

Once a picture is taken, custom software automatically reads the barcode and crops and straightens the image. It can't transcribe information from the labels, however, a task some consider to be the toughest remaining problem in mass digitization of museum specimens. Labels include, at a minimum, the name of the collector, the type of organism, and the date and place it was collected, and many bear extra details such as descriptions of the environment. Yet they are often handwritten, some in the archaic scripts of 18th and 19th century naturalists. What's worse, they don't follow standardized formats, so it's hard for a software program to automatically tell which part to put in which database field.

Currently, optical character recognition (OCR) programs do a decent job of turning printed labels into blocks of text, though handwriting still comes out as gibberish. Some institutions are using OCR to sort images of labels so that people can more easily transcribe them into databases, says Barbara Thiers, director of the herbarium at the New York Botanical Garden in New York City. For example, transcribers can search OCR outputs for the names of particular countries or collectors, then work on similar specimens in groups.

Some researchers have tackled the harder problem of automatically transferring OCR outputs to database fields using machine learning and natural language processing algorithms. At the California Academy of Sciences (CAS), which is digitizing its herbarium in San Francisco, such parsing software "helps speed up the entire process," says Anne Barber, a former CAS digitization project manager who helped develop it.

But human users still have to go through the output to correct errors, and many institutions find it cheaper and faster to enter data the old-fashioned way. Some use crowdsourcing initiatives to spread out the labor; NMNH has had a volunteer transcription program for years. But with the new conveyer belt setup, the NMNH herbarium is producing images too fast for volunteers to keep up. Instead, the pictures are sent to a company called Alembo in Suriname, where professional transcribers type in the label data by hand at a rate of about 60 specimens per hour, according to the company.

The digitization efforts are paying off. When Kellen Calinger, a forest ecologist at Ohio State University, Columbus, wanted to evaluate which native plants are likely to perish and which invasive ones might take over as the climate warms, she turned to 200,000 digitized specimens from the university's herbarium. Using the collection date and location on each label, Calinger assessed changes in the abundance and distribution of more than 200 plant species in Ohio over 115 years, and compared the results with historic temperature records.

Photographs of the specimens showed whether plants were in bloom, allowing her and her colleagues to conclude that the nonnative species most likely to expand their ranges were those that could adapt to rising temperatures by changing when they flowered. That, in turn, can suggest which invasive species are most threatening in a warming world and which natives are at greatest risk from the competition. "Having these predictive metrics can be really useful when selecting which species might be most important to focus our conservation efforts on," says Calinger, whose work was published last year in the journal *Biodiversity and Conservation*. Other researchers have used hundreds of thousands of digitized plants from museums across Australia to study nonnative species, examining when and where they first appeared on the continent to identify likely sources of future invasions.

**MOST MUSEUM SPECIMENS** are more challenging than plants, and each category presents its own set of problems. There is no automation yet for things stored in jars; to photograph a fish, someone must pluck the dripping specimen and its label out of alcohol and arrange them on a tray. Insects, the most abundant type of specimen worldwide, are a nightmare to digitize. A single drawer can hold hundreds of bugs on pins, sometimes so close their wings overlap. Their fragile bodies often hide the labels below. A careless touch snaps off legs.

Rather than photographing specimens such as insects, fossils, and shells one-by-one, some institutions are capturing images of a whole drawer-full at once. Invertnet, a collaboration that aims to digitize more than 50 million insects and other arthropods in collections across the midwestern United States, is employing a type of robot called BugEye. Developed at the University of Illinois, Urbana-Champaign (UIUC), it looks, appropriately enough, like a giant mechanical spider hanging upside-down over the drawer. BugEye's central camera moves over the drawer in a series of passes, taking hundreds of overlapping images that it stitches together into a high-definition composite. Many of the labels are hidden under the bugs, but human users reveal some of them by tilting the drawers, letting the robot work from different angles.

"With these whole-drawer images I can just sit at my desk and browse through all these collections virtually," says Christopher Dietrich of UIUC, an entomologist and Invertnet project leader. "If I see something that I've never seen before, I can contact the curator of that collection and say 'Hey, the specimen No. 10 in this drawer looks like it might be a new species.'"

Pinned insects like these endangered Taylor's checkerspots (*Euphydryas editha taylori*) are difficult to digitize, with fragile bodies often obscuring the labels below.

PHOTO: A. WARREN/MCGUIRE CENTER FOR LEPIDOPTERA AND BIODIVERSITY

Technicians can manually divide a whole-drawer image into images of individual bugs by drawing boxes around them with a cursor. But this is "pretty soul-destroying work," says Lawrence Hudson, a scientific software engineer at the Natural History Museum (NHM) in London. Last November, he and colleagues published a new open-source software package called Inselect that helps automate this process, defining images of individual specimens and streamlining the creation of database records. People still have to adjust the borders of the boxes and transcribe label data, Hudson says, but it speeds up the process; Dietrich says that Invertnet plans to incorporate it into their workflow in the coming months.

Even slower-paced digitization projects can benefit insect researchers. Male silver-spotted skipper butterflies (*Hesperia comma*) tend to grow larger in years that are warm when the caterpillars are feeding, according to a study published last February in the *Journal of Animal Ecology*. To reach that conclusion, Phillip Fenberg, an ecologist and evolutionary biologist at the University of Southampton in the United Kingdom, and colleagues measured the wings of 331 silver-spotted skippers collected over nearly a century. London's NHM recently photographed them one at a time for a digital archive.

The butterflies from Fenberg's study were among nearly half a million butterflies and moths NHM has digitized thus far—a small chunk of its 10-million-specimen Lepidoptera collection, notes Gordon Paterson, chair of the digitization project and a senior researcher in NHM's department of life sciences. As more specimens and more species are added, researchers will be able to assess larger trends, such as the interplay between climate, body size, and reproduction. Certain butterfly species are able to have multiple generations per year, whereas others are limited to one. Fenberg and colleagues suspect the latter butterflies respond to warming climates by increasing body size, whereas other species get smaller and put the saved energy into having generations faster. "I expect to see some amazing things happening with the data," Paterson says. "I think it's really important to have the museums' knowledge released out into the wild."

In the end, these diverse projects will do more than aid individual research projects. The digitizers' ultimate goal is to build an interconnected online library where everyone can see and study specimens stored all over the world. Such a network could allow both professional and citizen scientists to take full advantage of the data their forebears spent centuries collecting. "This is a democratization of knowledge and data," says Mark Lindeman, general director of Picturae, the Dutch digitization company in Heiloo that designed and helps operate the NMNH conveyer belt system. "I really believe that opening up this data will improve the knowledge we have of the world around us."

But the key to doing that may be developing even more innovative digitization solutions that save museums time and money. "The technology is changing so fast," Orli says. "What we're doing now, we'll probably be laughing about in 5 years." ■

Published by AAAS

**Museum drawers go digital**
Nala Rogers (May 12, 2016)
*Science* **352** (6287), 762-765. [doi: 10.1126/science.352.6287.762]

Editor's Summary