
A file cabinet at the Patuxent Wildlife Research Center in Maryland holds some of the center's six million bird-migration observation cards dating back to the late 1800s. The hand-written cards contain data about sightings of birds such as the ruby-throated hummingbird (right), often spotted in the 1930s when fruit trees bloomed in spring. Now being digitized, data from these cards will be stored on a U.S. Geological Survey database.

By Divya Abhat

*Divya Abhat is a science writer for The Wildlife Society.*

Credit: US Geological Survey

# Don't Hold That Thought

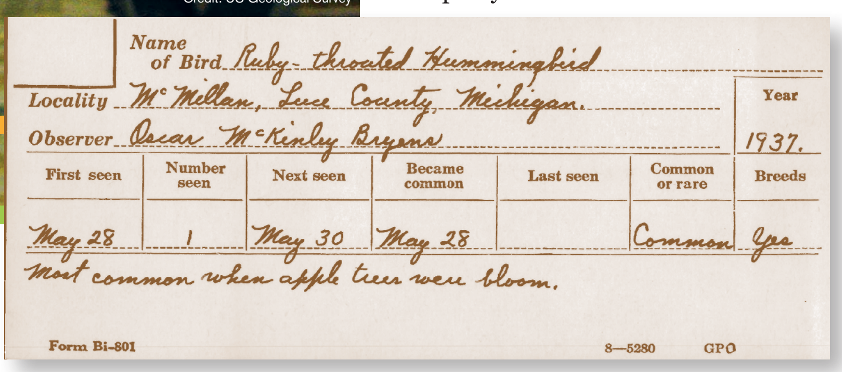
## *The benefits—and challenges—of efficient data sharing*

There's good news and bad news. The good: Since April 2009 the U.S. Geological Survey's (USGS) Patuxent Wildlife Research Center (PWRC) has recruited more than 1,500 volunteers worldwide to begin digitizing six million bird migration observation cards. The cards, stacked in 40 filing cabinets in the PWRC, date from the late 1800s to the 1960s. Once digitized, the information will be stored in a USGS North American Bird Phenology Program database along with other data on bird phenology across North America. As soon as the digital data are ready, the USGS will open the database for researchers, thereby increasing the pool of knowledge about migratory birds.

Now for the bad news. Around the world millions of similar handwritten wildlife observations scribbled over the centuries are also waiting to be digitized. When they are, they will be stored on any of hundreds of databases or web portals. Wildlife professionals may have access to more information, but may not have a clue where or how to find it. And that's just part of the problem.

### Data Overload

Wildlife professionals are well aware of the challenges that accompany the sharing and management of data. There's too much data—or not enough. Data are not presented in a standardized way and are not all openly accessible. Scientists aren't always sure where they should post their research or whether to make the data freely available. Those who do share data don't always update or archive them, and many choose not to include metadata (detailed information about a dataset that makes searching for data easier), leaving their work virtually inaccessible, incomplete, and inadequate for the rest of the scientific community. Clearly, data sharing and management vary in precision and accuracy as do the methods used to collect, store, and distribute data across the world (Van House 2002).



Name of Bird: Ruby-throated Hummingbird
Locality: McMillan, Luce County, Michigan.
Observer: Oscar McKinley Bryens
Year: 1937.

| First seen | Number seen | Next seen | Became common | Last seen | Common or rare | Breeds |
|---|---|---|---|---|---|---|
| May 28 | 1 | May 30 | May 28 | | Common | Yes |

Most common when apple trees were bloom.

Form Bi-801 8—5280 GPO

For the wildlife profession, this is a crucial issue. In these days of climate change and human sprawl, wildlife biologists are scrambling to study and share information about species biodiversity, genetics, disease, populations, and ecology. They're dealing with a broad range of datasets—on biology, geography, meteorology, geology—which can often contain politically and economically sensitive information. Yet access to such datasets is essential for effective conservation. "There are invasive species that come from one corner of the world and wreak havoc on another part of the world," says Vishwas Chavan, senior program officer for Digitization and Mobilization of Primary Biodiversity Data at the Global Biodiversity Information Facility (GBIF), a ground-breaking global infrastructure to facilitate free and open access to biodiversity data (see Tools and Technology, page 30). "Such a phenomenon cannot be studied with localized data and, therefore, the exchange and sharing of data are critical to understand the holistic picture of the state of biodiversity and model future trends."

Unfortunately that holistic picture has some holes, especially in countries such as Brazil and Indonesia, which have critical conservation issues but in some areas have limited access to the Internet. The people of Brazil's Iguazú Falls region, for instance, "have no GIS infrastructure, no research program, no remote sensing," says Ben White, NASA earth science system fellow at the University of Maryland. "But they've got the largest remaining forest fragment [in the world]." In addition, some developing countries may be unwilling to share information about their biological resources for fear that private companies will use it for individual gain. This situation is creating a widening information gap between the developing and developed world (Masood 2004). "The key issue is to make the best science accessible," says Falk Huettmann, associate professor at the Institute of Arctic Biology in Alaska. "If you really want to make an impact you have to have data available in China or India ... so that [researchers there] are able to make the right decisions."

"People at large are the true owners of the data and researchers are mere caretakers or custodians."

## That Forgotten Sandbox Lesson

Sharing doesn't always come easily. "There's a tradition in the science community that the data you collect belong to you rather than the public, even if the study is government funded," says Sue Haseltine, USGS associate director for biology. Understandably, research data may be considered the most valuable property of researchers, who are typically measured on their research outputs, so debates over data ownership, copyright, and control are commonplace (Council on Government Relations 2006). Survey results published in 2006 in *Academic Medicine*, for example, showed that 46 percent of life-sciences researchers were denied permission to use other scientists' published data (Vogeli *et al.* 2006). "This is irrespective of the fact that it is the taxpayers who fund the research work leading to the creation of data," Chavan says. "Thus, people at large are the true owners of the data, and researchers are mere caretakers or custodians of the data." Effective global conservation depends upon broadly shared standardized data. "Most of the big issues that confront wildlife management and wildlife science in this century are bigger than the people who are collecting individual bits of data," Haseltine says. "You cannot manage for climate change based on individual counties, or states, or projects."

Occasionally institutions and researchers will justifiably withhold sensitive data, such as the location and range of a threatened plant or animal, to protect the species from deliberate targeting or illegal trade (Chapman 2007). "If you're a state wildlife manager, you have very strict mission prohibitions about sharing individual locations of an animal," Haseltine says. Before opening data to the public, researchers must consider the threat and impact the information could have on a species. Most will then document the constraints for access and explain the need for and rationale behind withholding some information. On occasion researchers may simply delay data release. "We have the fear that when you put out caribou telemetry data, people hunt down the caribou," Huettmann says. The solution: "You just put a [two- to three-day] delay on it so the caribou have moved on."

Researchers may find little value in data that are extremely localized and area specific. Seth Wilson, program coordinator for Montana's Blackfoot Challenge Wildlife Committee, has been researching grizzly bear activity in the region since 2003. Over the years he has used satellite imagery and aerial photographs to map bear habitat; set up GPS locators for all local houses, trailheads, and beehives; and create grizzly bear habitat maps with additional information from the Montana Department of Fish, Wildlife, and Parks and the U.S. Fish and Wildlife Service. Wilson's data are available to all of his project partners, and he will regularly present his findings to area landowners and wildlife managers. However, all of Wilson's data are stored on his own machine. "This is a fine-scale dataset," he says, "and

Credit: New England Aquarium

Data transfer from cameras and clipboards to databases and datasets, though crucial, is oftentimes a painstaking process. Additionally, purchasing and maintaining hardware and software required for data and metadata entry can be prohibitively expensive.

it may not have relevance to someone outside of the area. However our data are available to the public upon request." Indeed, although data in isolation might not be universally relevant, when integrated, that information can open up new avenues for analyses, syntheses, and forecasting.

## The Devil in the Details

Metadata address the key elements that determine the usability and accessibility of a database. These data can include information about where, when, why, how, and by whom a study was done, as well as details about the focus of the study—whether it pertains to habitat of a species, population numbers, or migration patterns, for instance—and what software, tools, and technology were used. Metadata make targeted searches easier, and by having access to metadata, users can also determine whether a study will be of interest to them and their work.

Adding metadata can be a tedious and time-consuming task, however, which is why scientists and researchers will often opt out of doing it. In addition, purchasing and maintaining the hardware and software required for metadata files can sometimes be costly, making some managers wary of dealing with them (Digital Preservation Committee 2006). Yet because metadata enable speedy and accurate searches within a large database or set of databases, researchers consider the metadata one of the most crucial factors in data sharing and management. "It really comes home to the metadata," Huettmann says. "You can really only study data with metadata."

## A Moral Imperative?

Data sharing and management are crucial to any field of research, but the methods of sharing data vary depending on the field. In the biomedical world, for example, policies regarding data sharing and archiving are standardized and stringent. Several medical research publications and institutions require extensive and timely sharing of any biomedical data related to infectious diseases and human health. Because of that mandate, says Haseltine, "they have been able to develop systems to share this information more broadly." Based on the National Institutes of Health (NIH) 2003 Data Sharing Policy, for example, any NIH-funded research costing more than $500,000 must include an extensive plan to address the sharing and archiving of data (Final NIH Statement on Sharing Research Data). "I think there's more of a moral imperative to share data for people to be able to make the wisest health and treatment decisions, although a similar imperative should be applied to wildlife and ecosystem health," says Frank Biasi, director of National Geographic Maps' Conservation Projects.

Although some might say that policies related to data sharing and management in biodiversity aren't as standardized, others will argue that biodiversity is acting as a model for data sharing by other disciplines (Arzberger *et al.* 2004). "The global movement is in favor of data sharing," Huettmann says. "If you're not sharing you get peer-pressured into it." Several conservation-oriented journals, funding agencies, and institutions now encourage their contributors to adhere to guidelines that highlight the need to quickly make raw and published data widely and freely available while also safeguarding confidential and proprietary data such as location of a sensitive species. Recently, ZooKeys became the first open access, peer-reviewed systematics journal to launch simultaneous publication of primary data through GBIF, the Encyclopedia of Life, and other data access points (Penev *et al.* 2009).

Credit: David Safine/ USFWS

Biological technician Kate Soetaert jots down swan nest data at Scoter Camp Lake in the Yukon Flats National Wildlife Refuge. Occasionally, researchers will withhold sensitive data, such as nest locations or the range of a threatened plant or animal, to protect the species from harm.

Most data-sharing guidelines are not binding, however, so data sharing is left to the discretion of the author or researcher. To address this issue, the National Science Foundation (NSF) and the USGS are formulating stricter policies. The NSF-funded Long-Term Ecological Research (LTER) network, for example, requires LTER-funded data and information to be openly accessible online with "as few restriction as possible, on a nondiscriminatory basis." It also requires scientists to release information in a "timely fashion and with attention to accurate and complete metadata" (LTER Network Data Release Policy). Similarly, the USGS requires open access, metadata management, and public sharing of data and information to federal and nonfederal communities.

Credit: Becky Lettenberger

Nature's relics abound in the Patuxent Wildlife Research Center's lab where researchers can study data in print and online. Data, once collected, can be stored on any of several hundred databases and datasets with information about species biodiversity, genetics, disease, populations, and ecology.

In response to concerns that the United States has minimal investments in science and technology research, former President George W. Bush in August 2007 signed the America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Act (America COMPETES). The act calls for federal agencies to provide policy, procedures, and guidelines that will promote an easy and open exchange of data among agencies, the public, and policymakers. Federal research has always been accessible through the Freedom of Information Act (FOIA), which is based on the principle of openness in government and provides the public with access to federal agency records. This doesn't always apply to raw data, however. Additionally, FOIA does not apply to information at the state and provincial levels. More problematic, says Huettmann, "the FOIA is not digital. It just [applies to] hard copy."

## The Data Dilemma

It is ironic that the two chief complaints about biodiversity data are seemingly contradictory: that there are not enough being shared, and there are too many ways to share them. Numerous biodiversity databases, datasets, and web portals operate without comprehensive standardized methods of data sharing and management. This is complicated by the amount and complexity of geospatial data, which dominates biodiversity and conservation databases. "A lot of these efforts are being built and managed on a shoestring so there's not much budget or capacity to do marketing or outreach or technical support," Biasi says. This leaves vast amounts of data spread across the Internet with no comprehensive inventory that can index all the information in one place. "There's a life cycle to data management, all the way from the discovery and acquisition to delivery and archiving," says Mike Frame, research and development director at the USGS Center for Biological Informatics. "There's a whole chain that you have to consider, and each link of it occurs in different phases and requires different resources and sometimes different skill sets."

Numerous organizations and online interfaces such as NatureServe Explorer and MaNIS are attempting to address this issue. These resources are designed to provide biodiversity data that can range from detailed local data on plant and animal species to data on vertebrate specimens in museums around the word (see Tools and Technology, page 30). And while these interfaces have served that role well, they are only some of more than 500 resources created to provide comprehensive information on nomenclature, specimen-focused, or species-specific data (Biodiversity Information Standards). Any scientist in search of a comprehensive picture of a species and its habitat will therefore have to refer to many different resources. "What is needed is interoperability among all these resources, so that one can move easily between them," says James Edwards, executive director of the Encyclopedia of Life, who also notes that a recent international meeting, e-Biosphere 09, focused on this topic. Further, if a user is unaware of these resources or a researcher does not post his or her information online, the study will still remain hidden.

Research resulting in data that can be shared will always be the foundation of any scientific field, but given the challenges and controversy of data sharing, researchers may ask, "Why are we all doing this?" Huettmann's answer: "We want to have less wilderness lost and fewer species on the endangered species list. We want to make a real impact, and there needs to be better management that is traceable in real terms." ■

*This article has been reviewed by subject-matter experts.*

See this article online at *www.wildlife.org* for an abstract, bibliography, and additional resources.