



## NOT JUST FOR THE BIRDS

### Archiving Massive Data Sets

By Pam Frost Gorder

**A**MONG THE 7,000 BIRD SPECIES WHOSE SONGS ARE RECORDED AT THE CORNELL LABORATORY OF ORNITHOLOGY'S MACAULAY LIBRARY, THE PROTHONOTARY WARBLER

doesn't seem like a standout. It sings a simple tune—a one-note, “sweet, sweet, sweet, sweet, sweet.” But when Macaulay Library engineer Bob Grotke wants to test new audio equipment, this warbler's song is one of his top choices.

“It's very high-frequency, with rapid frequency sweeps—and very ‘bursty,’” he says, referring to the sharp way the bird punctuates every high-pitched tweet. He calls it, “a difficult little bird to capture accurately.”

Accuracy is paramount for Grotke and his colleagues, who are converting the world's largest animal recording collection from analog to digital to preserve as much information as possible for future scientists. Since 1999, they've digitized one-third of the collection and amassed 4 terabytes (Tbytes) of data. The sound collection dates back to 1929, and many of the earliest magnetic tape recordings have long since passed their expected lifetimes. Tapes are degrading—literally losing magnetic particles—with every play.

Macaulay Library engineers are engaged in a particularly dramatic race against time, but they aren't alone in their need to preserve massive amounts of information. From high-energy physics to climate science to biology, new instruments are gathering more experimental data that need to be retained for the long term. Meanwhile, other scientists need to retain results from huge computer simulations.

Right now, all around the world, data is being lost. To preserve its collections, the Macaulay Library is blazing a trail that others will have to follow.

#### From Byrd to the Birds

“Music is tough enough to record,” Grotke says. He should know: he once engineered a Tony Bennett album and recorded a host of jazz greats, including guitarist Charlie Byrd. In spite of this experience, he claims, “bird song is a

technical nightmare.” Animals communicate across broad frequency ranges—many with very complex vocal structures, some of which humans can't hear. Then there are “bursty” animals, such as the Prothonotary Warbler, which doesn't start singing low and quiet like many birds—it screams at the top of its lungs for the entire song. Human voices and musical instruments are tame by comparison. Grotke says he came to the Macaulay Library in part to confront this challenge, but also to preserve a precious resource.

“The presence or absence of birds in a given habitat is a well-known indicator of the health of that environment. Accurately preserving these sounds and making them available to future generations for research and global monitoring efforts is something I am very passionate about,” he says. When the Ivory-Billed Woodpecker—once thought to be extinct—was recently sighted in the Arkansas woods, scientists used 70-year-old recordings from the Macaulay Library to identify the bird's call.

In a recent issue of *The Auk* (vol. 122, no. 3, 2005), Macaulay Library engineers described the technical issues they face. The key to preserving the sounds is digitizing them at a high enough sample rate to accommodate the frequency range and then storing them so that all the information can be retrieved easily. Typical CD-quality sound is sampled at 44.1 kilohertz (kHz) with a 16-bit data stream, but many Macaulay Library tapes far exceed the bandwidth this format offers. Grotke opted for 96-kHz sampling for most birds, 192 kHz for bats and marine mammals, and a 24-bit data stream. He pieced together an analog-to-digital converter and a digital audio workstation that fully supported this unique data structure.

For data storage, the engineers needed a technology that would last—and something that future librarians could easily retransfer to new data systems. DVD-Audio seemed the natural choice, until the researchers realized the music industry's copy-protection technology would automatically downsample any sound to 48 kHz (16 bits) when they tried to play it back. Because the whole point of preservation is to make the full frequency range available for study, the engineers turned to DVD-ROMs, storing the high-resolution audio files as data. So far, they've filled three and a half of

# computing

in SCIENCE & ENGINEERING

---

## EDITOR IN CHIEF

**Norman Chonacky**  
cise-editor@aip.org

---

## ASSOCIATE EDITORS IN CHIEF

**Denis Donnelly, Siena College**  
donnelly@siena.edu

**Douglass E. Post, Carnegie Mellon University**  
post@ieee.org

**John Rundle, Univ. of California, Davis**  
rundle@physics.ucdavis.edu

**Francis Sullivan, IDA Ctr. for Computing Sciences**  
fran@super.org

---

## EDITORIAL BOARD MEMBERS

**Klaus-Jürgen Bathe, Mass. Inst. of Technology,**  
kjb@mit.edu

**Antony Beris, Univ. of Delaware, beris@che.udel.edu**

**Michael W. Berry, Univ. of Tennessee, berry@cs.utk.edu**

**George Cybenko, Dartmouth College,**  
gvc@dartmouth.edu

**Jack Dongarra, Univ. of Tennessee, dongarra@cs.utk.edu**

**Rudolf Eigenmann, Purdue Univ.,**  
eigenman@ecn.purdue.edu

**David Eisenbud, Mathematical Sciences Research Inst.,**  
de@msri.org

**William J. Feiereisen, Los Alamos Nat'l Lab,**  
ill@feiereisen.net

**Geoffrey Fox, Indiana State Univ.,**  
gcf@grids.ucs.indiana.edu

**Sharon Glotzer, Univ. of Michigan, sglotzer@umich.edu**

**Anthony C. Hearn, RAND, hearn@rand.org**

**Charles J. Holland, Darpa, charles.holland@darpa.mil**

**M.Y. Hussaini, Florida State Univ., myh@cse.fsu.edu**

**David P. Landau, Univ. of Georgia,**  
dlandau@hal.physast.uga.edu

**B. Vincent McKoy, California Inst. of Technology,**  
mckoy@its.caltech.edu

**Jill P. Mesirov, Whitehead/MIT Ctr. for Genome  
Research, mesirov@genome.wi.mit.edu**

**Charles Peskin, Courant Inst. of Mathematical Sciences,**  
peskin@cims.nyu.edu

**Constantine Polychronopoulos, Univ. of Illinois,**  
cdp@csrd.uiuc.edu

**William H. Press, Los Alamos Nat'l Lab., wpress@lanl.gov**

**John Rice, Purdue Univ., jrr@cs.purdue.edu**

**Ahmed Sameh, Purdue Univ., sameh@cs.purdue.edu**

**Henrik Schmidt, MIT, henrik@keel.mit.edu**

## NOT A DEAD ISSUE

**B**ob Grotke sees a need for new computer algorithms that compress massive audio and video data losslessly. At the Macaulay Library, the problem is one of sheer bandwidth and storage capacity, but other types of research have data sets with many variables and dimensions that must be “squashed” together for storage and then retrieved intact.

Raymond L. Orbach, director of the Office of Science at the US Department of Energy (DOE), was also at the Association for the Advancement of Science (AAAS) symposium on data collections. His office plans to initiate a long-term research program to address this so-called “curse of dimensionality.” As data sets have grown larger, researchers have grown frustrated, he says. Data mining is more cumbersome, meaning that important information might be missed.

During his presentation, Bruce Schatz of the University of Illinois, Urbana-Champaign, put it a bit differently. “This is a problem that hits researchers where they live,” he says, because “data is disappearing.” A professor of library and information science, Schatz co-leads the university’s effort to build an online information system called BeeSpace ([www.igb.uiuc.edu/beespace/](http://www.igb.uiuc.edu/beespace/)). The project will analyze

the library’s 12 DVD “jukeboxes,” each of which holds 480 disks. A similar effort to archive the library’s relatively new video collection takes up half as much disk space. Duplicates reside in a safe location off campus.

### Who Cooks for You?

The ultra-high-quality recordings are used only in-house, but the library will soon offer downsampled copies on the Web ([www.animalbehaviorarchive.org](http://www.animalbehaviorarchive.org)). Right now, users can listen to a few birds, frogs, and marine mammals, but the Web site will ultimately link all the information scientists will need to study these animals. Critical data include the locations in which recordings were made; maps of the animals’ habitats; sound, video, and graphical representations of song frequency called spectrograms; and even the mnemonics that field scientists memorize to help them recognize an animal’s call when they hear it. (The Prothonotary Warbler’s call is often translated as, “sweet, sweet, sweet, sweet, sweet,” but the mnemonic for the Barred Owl’s hoot is more typical of birds’—and birders’—creativity: “WHO cooks for YOOOU? WHO cooks for YOOOU-ALL?”)

Large data sets are often idiosyncratic, says Jeff Dozier, professor of snow hydrology, Earth system science, and remote sensing in the Donald Bren School of Environmental Science and Management at the University of California, Santa Barbara. Speaking at a symposium on

genes and behavior, using the Western Honey Bee as a guide. Scientists will compile a detailed database of gene expressions for hundreds of individual bees and link the genes to each bee's unique societal role.

For its part, the DOE plans to boost data storage funding from US\$34 million to US\$37.6 million for 2007, Orbach says. His office boasts 100 petabytes (Pbytes) of data storage, which he expects to more than double by 2009 to make room for burgeoning experimental and simulation data. Then comes the challenge of getting the data to the people who analyze them—ideally, in real time. "This is not a dead archival issue," he adds.

Anita Jones, the Lawrence R. Quarles Professor of Engineering and Applied Science at the University of Virginia, chaired a 2003 US National Science Board workshop on this issue. The resulting report, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," is available on the Web ([www.nsf.gov/pubs/2005/nsb0540/](http://www.nsf.gov/pubs/2005/nsb0540/)). At AAAS, she commented that the task of preserving large digital data sets is the purview of all science and engineering. "It may well be the best thing for science that NSF [US National Science Foundation] and other agencies make a long-term investment" in data collection, she says.

data preservation and management at the February 2006 meeting of the American Association for the Advancement of Science in St. Louis, Missouri, Dozier outlined some of the challenges from the data author's perspective. For his own work, he builds environmental models of snow accumulation and snow melt in the Sierra Nevada mountains using satellite data from NASA's Earth Observing System. He downloads 36 Mbytes of data per day—an amount that's easy enough to store on disk—but the challenge is to conveniently store the data in a way that other scientists can use for future research.

The way people access massive data sets is also changing, Dozier says. Rather than simply obtaining data from major data centers such as government agencies, researchers are beginning to share their own data products with each other via the Internet. In this scheme, the data's lineage becomes important. "If you use my data product, you want to know what went into it. In best practices, there would be an electronic version of a research notebook that preserves that information," he says.

An archive should also contain a description of the computations performed on the data so that others can reanalyze them, fill in missing information, or correct errors. Dozier uses a wrapper script that works passively in the background of his applications to store this information. His research group will have their archived data products available online soon at [www.snow.ucsb.edu](http://www.snow.ucsb.edu).

## DEPARTMENT EDITORS

**Book Reviews:** Bruce Boghosian, Tufts Univ., [bruce.boghosian@tufts.edu](mailto:bruce.boghosian@tufts.edu)

**Computing Prescriptions:** Isabel Beichl, Nat'l Inst. of Standards and Tech., [isabel.beichl@nist.gov](mailto:isabel.beichl@nist.gov), and Julian Noble, Univ. of Virginia, [jvn@virginia.edu](mailto:jvn@virginia.edu)

**Computer Simulations:** Muhammad Sahimi, University of Southern California, [moe@iran.usc.edu](mailto:moe@iran.usc.edu), and Dietrich Stauffer, Univ. of Köln, [stauffer@thp.uni-koeln.de](mailto:stauffer@thp.uni-koeln.de)

**Education:** David Winch, Kalamazoo College, [winch@TaosNet.com](mailto:winch@TaosNet.com)  
**Scientific Programming:** Konstantin Läufer, Loyola University, Chicago, [klauder@cs.luc.edu](mailto:klauder@cs.luc.edu), George K. Thiruvathukal, Loyola University, Chicago, [gkt@cs.luc.edu](mailto:gkt@cs.luc.edu), and Paul Dubois, [paul@pfdubois.com](mailto:paul@pfdubois.com)

**Technology Reviews:** James D. Myers, [jimmyers@ncsa.uiuc.edu](mailto:jimmyers@ncsa.uiuc.edu)  
**Visualization Corner:** Jim X. Chen, George Mason Univ., [jchen@cs.gmu.edu](mailto:jchen@cs.gmu.edu), and R. Bowen Loftin, Old Dominion Univ., [bloftin@odu.edu](mailto:bloftin@odu.edu)

**Your Homework Assignment:** Dianne P. O'Leary, Univ. of Maryland, [oleary@cs.umd.edu](mailto:oleary@cs.umd.edu)

## STAFF

**Senior Editor:** Jenny Ferrero, [jferrero@computer.org](mailto:jferrero@computer.org)

**Group Managing Editor:** Steve Woods

**Staff Editors:** Kathy Clark-Fisher, Rebecca L. Deuel, and Brandi Ortega

**Contributing Editors:** Cheryl Baltes and Joan Taylor

**Production Editor:** Monette Velasco

**Magazine Assistant:** Hazel Kosky, [cise@computer.org](mailto:cise@computer.org)

**Technical Illustrator:** Alex Torres

**Publisher:** Angela Burgess, [aburgess@computer.org](mailto:aburgess@computer.org)

**Associate Publisher:** Dick Price

**Advertising Coordinator:** Marian Anderson

**Marketing Manager:** Georgann Carter

**Business Development Manager:** Sandra Brown

## AIP STAFF

Jeff Bebee, Circulation Director, [jbebee@aip.org](mailto:jbebee@aip.org)

Charles Day, Editorial Liaison, [cday@aip.org](mailto:cday@aip.org)

## IEEE ANTENNAS AND PROPAGATION SOCIETY LIAISON

Don Wilton, Univ. of Houston, [wilton@uh.edu](mailto:wilton@uh.edu)

## IEEE SIGNAL PROCESSING SOCIETY LIAISON

Elias S. Manolakos, Northeastern Univ., [elias@neu.edu](mailto:elias@neu.edu)

## CS PUBLICATIONS BOARD

Jon Rokne (chair), Michael R. Blaha, Mark Christensen, Frank Ferrante, Roger U. Fujii, Phillip Laplante, Bill N. Schilit, Linda Shafer, Steven L. Tanimoto, Wenping Wang

## CS MAGAZINE OPERATIONS COMMITTEE

Bill N. Schilit (chair), Jean Bacon, Pradip Bose, Arnold (Jay) Bragg, Doris L. Carver, Kwang-Ting (Tim) Cheng, Norman Chonacky, George Cybenko, John C. Dill, Robert E. Filman, David A. Grier, Warren Harrison, James Hendler, Sethuraman (Panch) Panchanathan, Roy Want

## EDITORIAL OFFICE

COMPUTING in SCIENCE & ENGINEERING

10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720

phone +1 714 821 8380; fax +1 714 821 4010; [www.computer.org/cise/](http://www.computer.org/cise/)



IEEE Antennas & Propagation Society

IEEE

Signal Processing Society





## How to Reach CiSE

### Writers

For detailed information on submitting articles, write to [cise@computer.org](mailto:cise@computer.org) or visit [www.computer.org/cise/author.htm](http://www.computer.org/cise/author.htm).

### Letters to the Editors

Send letters to Jenny Ferrero, Lead Editor, [jferrero@computer.org](mailto:jferrero@computer.org). Provide an email address or daytime phone number with your letter.

### On the Web

Access [www.computer.org/cise/](http://www.computer.org/cise/) or <http://cise.aip.org> for information about CiSE.

### Subscribe

Visit [https://www.aip.org/forms/journal\\_catalog/order\\_form\\_fs.html](https://www.aip.org/forms/journal_catalog/order_form_fs.html) or [www.computer.org/subscribe/](http://www.computer.org/subscribe/).

### Subscription Change of Address (IEEE/CS)

Send change-of-address requests for magazine subscriptions to [address.change@ieee.org](mailto:address.change@ieee.org). Be sure to specify CiSE.

### Subscription Change of Address (AIP)

Send general subscription and refund inquiries to [subs@aip.org](mailto:subs@aip.org).

### Missing or Damaged Copies

Contact [membership@computer.org](mailto:membership@computer.org). For AIP subscribers, contact [kgentili@aip.org](mailto:kgentili@aip.org).

### Reprints of Articles

For price information or to order reprints, send email to [cise@computer.org](mailto:cise@computer.org) or fax +1 714 821 4010.

### Reprint Permission

Contact William Hagen, IEEE Copyrights and Trademarks Manager, at [copyrights@ieee.org](mailto:copyrights@ieee.org).

[www.computer.org/cise/](http://www.computer.org/cise/)

### That's No Siren

At the Borror Laboratory of Bioacoustics at Ohio State University, biology graduate student Miles Spathelf sits at a mixing table. He's digitizing reel-to-reel audio tape of sounds recorded in a Chinese rain forest. A shrill cry cuts through the forest backchatter, growing in pitch like an ambulance siren. "That's a gibbon," Spathelf says. Along with the National Sound Archive at the British Library, the Macaulay Library and the Borror Lab are the world's main repositories of animal sounds. All are digitizing their analog data.

Jill Soha, the Borror Lab's curator, points to a shelf of gold-coated CDs that comprise the lab's growing collection. As they do at the Macaulay Library, her staff keeps copies on an in-house server and stashes duplicate CDs off site for safety. They're also sharing downsampled versions of the sounds through a statewide educational network called OhioLink (<http://worlddmc.ohiolink.edu/media/borror/blbLogin/>).

Borror Lab director Doug Nelson coauthored the *Auk* paper with the Macaulay Library engineers. As computer memory gets cheaper, he sees bioacoustics labs moving toward

Join the IEEE Computer Society online at



[www.computer.org/join/](http://www.computer.org/join/)

Complete the online application and get

- immediate online access to **Computer**
- a free e-mail alias — **you@computer.org**
- free access to 100 online books on technology topics
- free access to more than 100 distance learning course titles
- access to the IEEE Computer Society Digital Library for only \$118

Read about all the benefits of joining the Society at

[www.computer.org/join/benefits.htm](http://www.computer.org/join/benefits.htm)


solid-state storage. With the advent of portable digital recorders, fewer scientists are dragging typewriter-sized reel-to-reel recorders into the field, and some data are coming to the lab already in digital form. "In the future, we'll probably just record straight onto hard drives and memory cards," he says.

Soha sits at a computer and plays one of more than 1,300 Song Sparrow calls from OhioLink. The mnemonic for this sprightly tune, Nelson says, is "maids, maids, put your TEA kettle-ttle ON!" Whereas Nelson studies subtle differences in birds' regional dialects, Soha listens for a change in the song—something to indicate what was happening to the bird at that moment.

"You hear that? Right now he's singing a different one," she says a minute later. The bird probably uses the same call to attract a mate and ward off rivals. Other calls could signal a food sighting or the swoop of a predatory Cooper's Hawk from above.

Nelson, Soha, and their colleagues are assembling a kind of avian sociology, and their work gets to the heart of theories in animal cognition. They want to know how birds learn their songs, and whether females prefer males with local accents.

The Borror staff spent three years digitizing their analog tapes; now they want to help smaller labs digitize their collections, too. That's something

Grotke would like to do as well. With two-thirds of the Macaulay collection still unarchived, and thousands of reels aging on shelves around the world, he ventures that his staff could stay busy for decades. One problem, though: the manufacturers of tape equipment are retiring the technology, and spare parts will soon be hard to come by. "I suspect that we won't be able to keep our hardware going for much more than 10 years," Grotke says. "So we've got a lot of work to do." 

**Pam Frost Gorder** is a freelance science writer and an associate editor of research communications at Ohio State University. She is based in Columbus, Ohio.

# UCES 2006



## Undergraduate Computational Engineering and Sciences Award Program

The UCES 2006 Award Program promotes and enhances undergraduate education in computational engineering and science. UCES Awards will be given to educators who have created resources or programs that enhance undergraduate CES education. Awards include a \$500 cash prize, a certificate and travel to the awards luncheon in Tampa, Florida, November 2006.

Apply Today!

## For additional information on the UCES 2006 Awards Program

including eligibility and how to apply, see the program website:

[www.krellinst.org/uces2006](http://www.krellinst.org/uces2006)

Application Deadline: June 15, 2006.

