



DIGITAL LIBRARIES COME OF AGE

By Pam Frost Gorder

A PERSONAL LIBRARY THAT FITS IN YOUR POCKET—IT WILL SOON EXIST, IF MICHAEL HART GETS HIS WAY. HART HEADS PROJECT GUTENBERG (PG), THE INTERNET'S oldest digital library. On 4 July 2006, PG turned 35. Digital libraries in academia and commercial publishing are coming of age right along with it, driven by the latest technology. But to Hart and others involved in digital libraries, the technology is the least interesting part of the story.

The Beginning

Hart created the first PG e-book in 1971, when some friends in a computer lab at the University of Illinois gave him free computer time. To create a document of lasting value, he uploaded the text of the US Declaration of Independence. Now some 20,000 books can be found at <http://gutenberg.org>—everything from *Alice in Wonderland* to *Zen and the Art of the Internet*. Another 80,000 can be found at PG's mirror sites around the world. All books in the project are in the public domain, and most are provided by a devoted cadre of volunteers.

Like typical digital libraries, PG contains two kinds of collections: hard-copy materials (such as books) that have been scanned and converted to plain text via optical character recognition (OCR) software, and newer documents that started out in digital form. Both can be linked by subject, keyword, or meta-information. Images can be tagged with keywords to enable searching in an otherwise plain-text environment.

As PG expands its collection, large brick-and-mortar libraries are trying to digitize their own collections. It's a major enterprise, one that Internet giant Google has taken on. In 2004, Google began a pilot project with the University of Michigan, Harvard University, Stanford University, Oxford University, and the New York Public Library to digitize their collections and make them searchable. Although that project is now the subject of litigation—Google is being sued by the Authors Guild and several publishers for

copyright infringement—it's nonetheless pushing digital library technology forward.

Unfortunately, scanning a book can sometimes mean damaging or destroying it. The binding warps text near the crease—what's called *guttering*—and warped text is hard for OCR programs to read. The result: a text file with garbled words at the beginning or end of every line. One solution is to press the book flat on top of a traditional scanner and crush the binding; another is to rip the book apart and feed it through a scanner page by page. But for rare books, neither option works.

Some desktop scanners have an angled surface, so the spine rests on a peak and one page at a time scans smoothly. Google has taken the care of its borrowed books further, by using a customized scanning machine with a robot hand that delicately turns the pages. Now a University of Kentucky project has yielded a method for scanning books that are too old or damaged to even open (see the "How to Scan a Book [Without Opening It]" sidebar).

Once a book is converted to text, it must be proofread. PG handles this step through a scheme called "distributed proofreading." Volunteers log into the site to proofread an uploaded book; they can read one page or the whole book—whatever they have time for. After one person proofs a page, a second person proofs the proof before it enters the archive.

Most of its books are stored in plain ASCII format. Illustrations from books, including artwork, scientific diagrams, and formulae, are saved as separate image files. Recently, PG began incorporating XML formatting into some documents, so that different reading devices can reformat the text. A text e-book can then become a Web page or a PDF file, with details like italics or images included.

Hart says that the beauty of PG is in its versatility. "I dare you to find a computer-and-program combination that won't read our books. I would just as soon read an e-book on a computer that was 20 years old as one that has all the latest bells and whistles. And I'm perfectly happy reading them on PDAs and PPCs [pocket PCs]."

For e-books to catch on commercially, people will have to adopt a convenient, affordable way of reading them. Portable game players are one possibility; they have an ad-

HOW TO SCAN A BOOK (WITHOUT OPENING IT)

Brent Seales and his colleagues at the University of Kentucky recently read the inside of an Egyptian scroll—without unrolling it. Their technique involves x-raying a closed document at different depths, reading the x-ray signature of chemicals in the ink, and using software to flatten out curved pages and recognize characters.

Since then, they've discovered a Hebrew manuscript fragment layered inside a book binding in the University of Michigan Library using a custom computerized tomography (CT) scanner made by physicist Joe Gray of Iowa State University. "This experiment is a breakthrough in that we're using data from a real fragment to show the viability of the technology," Seales wrote in a July 2006 email. "We've been given permission by the library to have a conservator physically remove the first layer to reveal the text below, so that we can compare that with our results. We'll be doing that later this summer."

He says improvements in resolution will drive this kind of imaging. Getting down to finer than millimeter-level resolution, which is currently the standard in health care, will be important for seeing text and carving in great detail. Seales also sees an opportunity for using immersive display environments. "Museums and libraries could build rooms for displaying their digital collections that are flexible, high-resolution, and very, very compelling," he says.

Ultimately, Seales would like to see his work encourage further archeological excavation because his scanning technique reduces the costs of analyzing new artifacts. "It is a bit like a digital treasure hunt—who knows what's out there?" he says. He offers videos of some of the team's projects at <http://halsted.vis.uky.edu>.

equate resolution for reading text. But Hart thinks that the "next big thing" for e-books will be a lot smaller than a Game Boy. "In the next year, there will be a billion new cell phones made," he says. "There will only be 100 million computers made. So, what are people most likely to have in their pocket?" He envisions a future in which people carry personal libraries with them, all the time.

IEEE Community

That vision isn't so different from the one that Ed Fox had in 1970, during his undergraduate days at the Massachusetts Institute of Technology (MIT). His advisor, Internet pioneer J.C.R. Licklider had just published the book *Libraries of the Future* (MIT Press, 1965), which predicted—quite prophetically—that libraries would one day exist as electronic repositories shared by an online community. Fox is now director of the Digital Library Research Laboratory

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

Bruce Boghosian, Tufts Univ., bruce.boghosian@tufts.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

Steven Gottlieb, Indiana University, sg@indiana.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

Rachel Kuske, Univ. of British Columbia, rachel@math.ubc.ca

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

Greg Wilson, University of Toronto,
gvwilson@third-bit.com

CONTRIBUTING EDITORS

Francis Sullivan, fran@super.org

Paul F. Dubois, paul@pfdubois.com

ANALOG ADVANCES

By Pam Frost Gorder

Paul Hasler does cutting-edge research that engineers 40 years ago would have found eerily familiar: in an increasingly digital world, he advocates processing data in analog form. Hasler is revisiting this “old” technology in a new way that could lead to more portable, energy-efficient devices—especially sensors.

The problem is this, though: we do our computing in digital, but life happens in analog. Physical quantities such as temperature, pressure, and acidity vary across a continuum. Early circuits were analog in that they were constructed as electrical analogies to these continuous systems; as the quantity being studied increased, current output through the circuit increased. The advent of digital circuits brought analog-to-digital converters, but also some new obstacles that might only be overcome with an innovative return to analog.

Hasler says he didn’t start out as an analog designer. He credits most of his inspiration to graduate study at Caltech with Carver Mead, a pioneer in silicon technology and neural networking. As a result, Hasler takes some of his cues from very efficient biological computation systems, such as the human brain.

CISE: Your concept is to create analog circuits that can be used to advantage in signal analysis practice. Can you describe a larger context into which this concept fits? Why go against the conventional wisdom of signal processing, which says “digitize first?”

Paul Hasler: There are a couple of reasons not to digitize first—power dissipation, for one. In the past, power wasn’t such a big issue because most computing was done on desktop machines. Now we’re all carrying laptops and iPods and so on. And analog computing can be quite a bit more power efficient.

For instance, if I need to perform 10,000 computations, that may cost me a milliwatt of power if I do it in digital. But in analog it would only require a microwatt. It’s the difference

between running a device on one AA battery for a week versus a year. In an analog system, you could lose more power to the battery’s parasitic losses than from running the device!

The second reason to process data in analog is, once you’ve gathered all your sensor data, can you process it fast enough? It’s a huge problem. CMOS [the complimentary metal oxide semiconductor] keeps getting better and better. The number of circuits doubles every 18 months, pretty much as Moore’s law predicted. But analog-to-digital converters only gain one additional bit of linearity every six years. The converters have not been able to keep up with the sensors, which are analog. So it makes sense to find a way to process the data in analog form first, before it gets converted.

CISE: Given that you will eventually have devices that are widely available, what are the particular benefits that make your approach to signal processing competitive, or even compelling, to employ?

Hasler: In analog, you can make devices that are smaller. You can use four analog transistors to do a computation that in the digital realm would require 1,000 transistors. Four transistors require a lot less room, even if each of the four is bigger than a single digital transistor.

Some high-end computing may be more practical when done in analog, too. When you need a huge number of computations, you can have many work units running in parallel or keep them all on one chip. And communicating between chips and boards is getting more and more power expensive. So the more units you can keep local the better.

One everyday device that could benefit from analog signal processing is the hearing aid. People want a hearing aid to be small, with long battery life and good sound quality.

CISE: How does your concept mesh with current practice, or how does current practice need to change to take advantage of what you have to offer? Are you asking for a paradigm shift in the way signal processing is done?

at Virginia Tech, and Chair of the IEEE Technical Committee on Digital Libraries (<http://ieeetcdl.org>), which formed in 1997.

Over the years, membership in the committee has remained strong, he says, and the annual ACM/IEEE Joint Conference on Digital Libraries (www.jcdl.org) is always well attended. Recently he’s noticed how the doctoral consortium has grown—more students are earning their PhDs in digital library science and engineering. “In the beginning, the meetings were about defining what digital libraries are,” Fox says, “and now we have doctoral students presenting research results and getting advice on the future.”

As to the technical challenges of setting up digital libraries, Fox points to several open-source content management systems that now make it easier, in particular DSpace (<http://dspace.org>), which the MIT libraries and Hewlett-Packard developed, and Fedora (www.fedora.info), which Cornell University Information Science and the University of Virginia Library developed.

Academic Matters

Of course, e-books aren’t the only library materials worth preserving: any respectable digital library needs scholarly journals, too. Journal Storage (JSTOR; www.jstor.org), the

Hasler: It's more a shift in the way we think about analog circuits, towards analog system design. There's no reason that analog can't go through the same evolution digital did. Today, we've got FPGAs [field-programmable gate arrays] with millions of digital gates. Why can't we do the same thing in analog?

Now, there are different design constraints and issues you have to deal with in both media, but we're starting to talk about analog as a systems technology much the way digital has been for many years. I think that's where the fundamental paradigm shift is coming.

As to meshing with current practice, we need analog systems that have all the programmability and configurability that digital systems do. We've spent a great deal of time on this problem, and we've built systems with 100,000 analog programmable parameters—what we'd call an FPAA, a field-programmable analog array. It's meant to be like an FPGA.

Of course, we still need digital computation, particularly for control and symbolic processing. So which parts of a system should be analog, and which should be digital? We need to build the analog side into the existing framework. The end product would be analog systems with classically digital—and creatively digital—circuits around them.

CISE: What are some problems you envision in getting instrument designers to change by adopting your paradigm for circuit design?

Hasler: Of the hurdles we see right now, I think most of them come down to what I'd loosely call education. It's teaching a classroom of people about how you do analog system design. But also, how do I explain the problem in a way that makes sense for a customer to use these analog systems?

Further, how do we build tools that help engineers do this efficiently? On the digital side, the tools are taken for granted. For FPGAs, a well-known set of tools lets you write the code and compile it. You don't even need to know the details of the circuit. We don't have that kind of tool set on the analog side. We're working it out pretty

rapidly, but that's going to be the most critical concern going forward.

CISE: What's in store for the future for your work in this area?

Hasler: We are looking at applications in speech recognition, image recognition, and compression. We are looking at systems that are not only programmable and configurable but that also start to be adaptive. These would be devices with very primitive neural systems that can learn what's going on in the world around them. It means having more cognitive behavior in our computing systems. I think we will learn more about the human brain and, at a very low level, make our machines a little more user friendly.

CISE: So does the human brain "compute" in analog? Is that even a reasonable question?

Hasler: It's a very reasonable question. I think most of the brain's computation is analog, though I also think that there is a lot of debate around this question. We are still very early in our understanding of neuroscience. One of the things I like about what we do with circuits is that we can make analogies to things we see in neurobiology. Sometimes that leads us to ideas about new experiments that neuroscientists can do.

MORE ABOUT PAUL HASLER

Paul Hasler received his BSE and MS degrees in electrical engineering from Arizona State University. He has a PhD in computation and neural systems from the California Institute of Technology. Hasler is an assistant professor at the Georgia Institute of Technology's School of Electrical and Computer Engineering, where he founded the Integrated Computational Electronics (ICE) laboratory, affiliated with the Laboratories for Neural Engineering. Atlanta is the coldest climate in which he has lived.

nonprofit scholarly journal archive, carries nearly 600 titles, the oldest of which dates back to 1665. Subscribers can read files in TIFF, PDF, or PostScript format.

Archives of a different sort can be found at the US National Science Digital Library (NSDL; <http://nsdl.org>). Created by the US National Science Foundation (NSF), the NSDL compiles K–16 educational resources from NSF-funded projects, educational Web sites, and other digital libraries. Through a new NSF project, computer scientists at Virginia Tech and Villanova University are building a user interface to connect the NSDL to college-course Web sites.

HighWire Press (<http://highwire.stanford.edu>), a division of the Stanford University libraries, boasts the largest repository of free, full-text, peer-reviewed content online. Citations are hyperlinked, and users can register to receive email alerts when a paper on a particular topic appears.

Connecting the Dots

Michael Keller, director of the Stanford University libraries, says that the financial cost of creating a digital library can be substantial. He expects that Stanford will have to allocate 1.5 Pbytes to store the books that Google is digitizing, and estimates that original digital content created by university fac-

DEPARTMENT EDITORS

Book Reviews: Mario Belloni, Davidson College, mabelloni@davidson.edu

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

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

Education: David Winch, Kalamazoo College, winch@TaosNet.com
News: Rubin Landau, Oregon State Univ., rubin@physics.oregonstate.edu

Scientific Programming: Konstantin Läufer, Loyola University, Chicago, klauffer@cs.luc.edu, and George K. Thiruvathukal, Loyola University, Chicago, gkt@cs.luc.edu

Technologies: James D. Myers, jimmyers@ncsa.uiuc.edu

Visualization Corner: Jim X. Chen, George Mason Univ., jchen@cs.gmu.edu, and R. Bowen Loftin, Texas A&M University, Galveston, loftin@tamug.edu

Your Homework Assignment: Dianne P. O'Leary, Univ. of Maryland, oleary@cs.umd.edu

STAFF

Senior Editor: Jenny Ferrero, jferrero@computer.org

Group Managing Editor: Steve Woods

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

Contributing Editor: Cheryl Baltes and Joan Taylor

Production Editor: Monette Velasco

Magazine Assistant: Hazel Kosky, cise@computer.org

Technical Illustrator: Alex Torres

Publisher: Angela Burgess, aburgess@computer.org

Associate Publisher: Dick Price

Advertising Coordinator: Marian Anderson

Marketing Manager: Georgann Carter

Business Development Manager: Sandra Brown

AIP STAFF

Circulation Director: Jeff Bebee, jbebee@aip.org

Editorial Liaison: Charles Day, cday@aip.org

IEEE ANTENNAS AND
PROPAGATION SOCIETY LIAISON

Don Wilton, Univ. of Houston, wilton@uh.edu

IEEE SIGNAL PROCESSING SOCIETY LIAISON

Elias S. Manolakos, Northeastern Univ., 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/



IEEE Antennas &
Propagation Society

IEEE

Signal Processing Society



ulty and students could fill another 100 Tbytes. The challenge, he says, is for universities to afford that much storage and set it up so that one copy of all documents is kept inviolate while another copy becomes a working file accessible by the readership. Such an undertaking also requires enough working memory and CPUs for digital archivists to layer new services above all that content. He calculates that licensing and subscription costs could easily run US\$7.5 million per year, and staff, equipment, and other expenses could tally up to \$3.5 million per year.

Despite the costs, Keller says that digital repositories have a big payoff for their host institutions. When Stanford digitized its card catalog, circulation of hard-copy materials went up by 50 percent. “That means that students and faculty use the collection 50 percent more because they can search, even in the metainformation about those books, and discover more works of relevance to them,” Keller says. “That’s a gigantic return on investment.”

The legal issues are another hurdle, and he acknowledges that digital libraries will have to work out the appropriate copyright permissions before they can attain their full promise.

Keller says that the life sciences and medicine disciplines have already benefited from online repositories that “connect the dots” buried in masses of text. “I think the same thing will be true in linguistics, anthropology, archeology, and literature,” he says. “I think as we look across the arts—music, dance, drama—and across cultures and languages, we’re going to see more of these dots being connected. And all that will happen because we can analyze the text across different languages and character sets.”

Hart points out that PG already carries some classical MP3s and sheet music. He sees no reason why digital archives can’t support all the arts. In fact, he has just returned from a visit to MIT’s Fab Lab, where engineers wowed him with 3D printouts—some carved with water jets or lasers, others fabricated from plastic—and now he’s thinking about digitized sculptures. “I have printouts of human hands that are so finely detailed that a palm reader could read them,” Hart says. “This is a direction for the future. Why should we stop with books? Why should we stop with paintings and pictures? You could print Michelangelo’s *David*, and Donatello’s *David* if you like, and compare them. You could print out every statue in the world.”

CS
SE

Pam Frost Gorder is a freelance science writer based in Columbus, Ohio.