A Conversation with Brewster Kahle

Creating a library

OF ALEXANDRIA FOR
THE DIGITAL AGE

Imagine—all the world's information at your service with just a few clicks of the mouse. It's a dream that Brewster Kahle has held onto for the past 20 years and is now seeing through to reality in his role at the Internet Archive, where he serves as chairman of the board. The Internet Archive was founded in 1996 to build an "Internet library" that will offer permanent access for researchers and scholars to historical collections that exist in digital format. Kahle is the force behind that effort.

Prior to his work with the Internet Archive, Kahle pioneered the Internet's first publishing system, known as WAIS (Wide Area Information Server), which was sold to AOL in 1995. He then cofounded Alexa Internet, which was sold to Amazon.com in 1999. Kahle earned a B.S. from the Massachusetts Institute of Technology in 1982. He studied artificial intelligence with Marvin Minsky and W. Daniel Hillis. In 1983, he helped start Thinking Machines, a parallel supercomputer maker, serving as a lead engineer for six years.

Discussing the potential of the Internet Archive with Kahle is Stuart Feldman, vice president of Internet technology for IBM. Before that, he was director of the IBM Institute for Advanced Commerce and head of computer science research.

Prior to coming to IBM in 1995, Feldman spent 11 years at Bellcore, where he held several research management positions. He spent 10 years before that as a computer science researcher at Bell Labs. Feldman was a member of the original Unix research team and is best known as the creator of the Make configuration management system, as well as the author of the first Fortran-77 compiler.

Feldman received an A.B. in astrophysical sciences from Princeton University and a Ph.D. in applied mathematics from the Massachusetts Institute of Technology.

STUART FELDMAN: How is it that you ended up in this most amazing role as the digital librarian of the Internet Archive? You had a string of obvious successes, making a major mark on a number of companies. Then you made this interesting apparent left turn into running a unique nonprofit specialized service.

BREWSTER KAHLE: This is all part of one theme that was floating in the air when I was in college: to build a digital library. The

thing that gets me springing out of bed in the morning and has for the last 20 years is the idea that we could have universal access to all knowledge.

It goes back very deep in the human psyche to the

PHOTOGRAPHY BY TOM UPTON

Library of Alexandria, which was in many ways the culmination of the Greeks' vision of knowledge as being worthwhile in and of itself. The idea is to take the Library of Alexandria another step further and make the published works of humankind accessible to everyone, no matter where they are in the world. We hope that then everyone can add to this grand library. Current computers and the Internet are making this conceivable. This seems to be the opportunity of our time, in the way that the generation before got to lay claim to landing a man on the moon. That was something that humankind can point at for centuries as a worthwhile achievement.

SF: What do you picture as the content? You referred to the published literature. Then you're obviously talking about being able to add video literature or radio literature.

BK: Humankind started recording things with the Sumerian tablet, so we might as well start there. We're talking about all books, all music, all video, all Web content, all software ever produced that was meant for any form of dissemination or for passing down from one generation to the next. It's not necessarily everybody's musings inside their heads. We'll cut our area into a smaller, more manageable set.

SF: So not everybody's laundry bills will necessarily be included.

BK: I don't think that's the first order of business.

SF: This is ambitious enough.

BK: Yes, but it's also quite doable. There are four questions: Should we do this? Can we do this? May we do this? And will we do this?

The first question of *should* we do this, I'm going to take as almost a postulate of yes.

SF: Because, obviously, not enough people have taken that as a postulate, since it wasn't being done very effectively before you.

BK: Yes, it's baked into the Enlightenment era of humankind—that knowledge is important to fulfilling ourselves as people and for building societies that grow and prosper. It's also baked into the American Constitution. It's fairly fundamental to the Renaissance, which is the rebirth of the Greek ideals.

I've grown up within this idea that universal education is good, and that people, if they can build on the works of others, achieve more. But this approach is not always in favor. Not all times in history encourage open societies and open knowledge.

SF: Your statements sound very American.

BK: Absolutely, I'm very American. I see what we're doing as being very much in the tradition of Ben Franklin's and Carnegie's vision of the library system and sort of the

Thomas Jefferson ideal of making an educated populace.

Then there is the question of "can we?" Within technological audiences, this is often the issue.

The "may we?" question is legal and societal.

SF: You're doing it, so obviously there is a way. When did you decide you could do this?

BK: While going to a technical college in the '70s, it became quite clear with the advent of Moore's law that you could name the year when all books, all movies, all music could be stored on computers.

SF: Presumably, things like the Internet added a new wrinkle here.

BK: How to move around all this information was a piece we were missing, and that's why many of us worked on an open Internet. The storage looked like that would all be taken care of. And the computation, no problem.

Let's consider the question of how much information there is. If you break it down, it turns out to be not that big of a deal. The largest print library in the world, which is the Library of Congress, has about 28 million volumes. A book is about a megabyte. That's just the ASCII of a book, if you put it in Microsoft Word. So 28 million

megabytes is 28 terabytes, which fits in a bookshelf and costs about \$60,000 right now. Storing books in ASCII is no problem, and the scanned images are more but still affordable.

Scanning books costs between \$5 and \$20. That's the mechanical cost if you just wanted to scan a book and end up with the images of the pages at high enough resolution that you could print it on a high-end laser printer so it would be a good facsimile at 600 DPI, color—a nice-looking book. So books are doable, in terms of technology.

Now let's take music. It's been estimated that there are about 2 to 3 million albums. In terms of salable units—

things that were sold as either 78s, LPs, or CDs—that's the universe of commercial music. If you do the math again, it's a few more of your bookshelves. So you're still not talking about anything daunting.

If you take movies and video, Rick Prelinger [founder of a film collection known as the Prelinger Archives] estimated that the total number of theatrical releases of movies was between 100,000 and 200,000. Again if you do the math, based on DVD quality, you come up with low numbers of petabytes [one petabyte is 1 million gigabytes].

SF: So, across a society this is not a big deal.

BK: Correct, we can afford this. The cumulative budgets of all of the libraries in the United States has been estimated between \$12 billion and \$24 billion a year. Interestingly, between one-quarter and one-third of that money (\$3 to \$8 billion) now goes to publishers' products. That's a lot of money, and everyone gets a lot out of it. With new technology we can multiply the effect of our spending in terms of serving the public and rewarding creators.

SF: Are people going to read books on their computers?

BK: For delivering public-domain books, the idea of read-

ing on screens is still far from ideal. We developed a way not only to combat the need to have a computer to read a book in our archive, but also to let people read them the old-fashioned way: the Internet bookmobile. Our general philosophy is to use commodity components, so we build a bookmobile that costs a total of \$15,000 including the car.

SF: This is a bookmobile without any books?

BK: Without any physical books. It prints them on demand. There is a satellite dish on top, a printer, a binder, and a cutter, and you walk away with a paperback of any of the public-domain books available on the 'Net.

SF: What's the incremental cost for a typical book?

BK: A 100-page black-and-white book with current toner and paper costs in the United States is \$1, not figuring labor costs, rights costs, or depreciation of capital. That's an interesting number, because at a buck a book, it turns out that for a library, it could be less expensive to give books away than to loan them. In his book, *Practical Digital Libraries*, Michael Lesk reported that it cost Harvard incrementally \$2 to loan a book out and bring it back and put it on the shelf. This is not figuring in the warehousing

costs and all the building costs. This is just the incremental cost of loaning a book out.

Even if you put some fee in for the author, it looks cost effective to print and bind many books locally.

SF: So, running a self-service kiosk would be...

BK: ...more cost effective.

SF: And you could let people burn them afterward if they dare. The book would say, "Please do not return this book."

BK: Or, "Please give it to somebody else." I think we're not setting ourselves up for tearing down forests by this system since people may be more likely to read books they have worked to print and bind. We are trying to avoid the inefficiencies of the current book distribution system, and at the same time offer a much broader range of books to everyone.

A year ago in San Francisco, we developed a print-on-demand bookmobile. I drove it across the country with my 8-year-old son, making books at schools, libraries, museums, and even in front of the Supreme Court. It worked.

We have now spun off a not-for-profit called Anywhere Books that's pursuing this idea. World Bank has funded a test that I was delighted to help launch in Uganda recently. If we could make this technology work in San Francisco and in rural Uganda, then we might have something.

SF: What has the demand been like?

BK: They love it. In this rural area they have created a reading program for the first time. This is the first time some of these kids have ever owned a book. We would like to see this grow within the library system and the Internet café system.

SF: Ignoring the vehicle, what does it cost to run?

BK: The capital cost is about \$5,000 or \$6,000 in the United States to buy the printer and binder and stuff. We think interested companies can get this below \$2,000, including the computer, with some creative product design. At that point you could have tens of thousands of these very quickly.

SF: This really sounds wonderful. Now, what are the flies in the ointment? Why don't I see a truck going up and down the street right now?

BK: The technology is still quite early. But I'd say the biggest barrier for achieving this goal is not technological. The technology problems are easy. Where we find the stumbling block now is actually the mind-set change that this is possible.

I'd say we have the three characteristics required to be able to pull this off, and they're in our hands for the first time in history.

We have the storage technology to be able to store all knowledge again.

We have the mechanism of doing distribution—universal access—using the Internet for getting things close to people. And then we need different mechanisms for the last mile.

The third characteristic, which is probably the least appreciated, is that we have the political will and the societal will.

With those three—the storage, the distribution, and the political will—we can leave something for our children that we can be proud of.

SF: Does that mean a permanent establishment of some sort?

BK: Yes, but I would say we need more than one establishment. If you look at the history of libraries, you see that they tend to be burned. The new guys don't want the old stuff around. So the lesson of the first Library of Alexandria is "don't have just one copy."

The collections of the Internet Archive are here in San Francisco, where we're very conscious of being in an earthquake zone. Also, we're in an upstart country that's only 200 years old. All sorts of things can go wrong.

Some scientists have learned to have copies of seed banks and data sets on different continents to aid preservation. We, in the library world, could do the same sorts of things.

SF: Would those hubs be complete mirrors?

BK: We envision complete copies of everything else in other Internet Archives around the world. They may have limited rights of what they can do with them. But at least it's preserved.

Our first agreement in this direction is with the new Library of Alexandria in Alexandria, Egypt, where we donated a copy of our collection in 2001. We have donated 100 terabytes of computer facilities and the data to go on it. Raj Reddy from Carnegie Mellon University has donated book-scanning facilities, so the library is digitizing its Arabic collection.

SF: It's the upgraded version of the Library of Congress.

BK: I would not go that far, because the collections in the Library of Congress are fantastic, but we hope that some of these ideas will be widely adopted.

But the key thing now is access. The way to start in this game is with a petabyte, a gigabit per second, and \$100 million in endowment.

SF: And how much computing?

BK: The computing that comes along with a petabyte—a thousand or a couple of thousand computers.

SF: So you need 1,000 processors, aggregate external networking of a gigabit a second, and a petabyte of persistent storage.

BK: Currently that is what it takes, but it will be beyond that soon. The endowment of say, \$100 million, can provide a funding stream to keep the bits accessible and fund the transition to new technologies as they come along.

We now have San Francisco and Alexandria getting there. The next one we hope will be in Amsterdam, because it's a really good place for bandwidth, technology, and a cultural in-between for Europe.

SF: What about the intellectual property law issues? What about control issues? What about the export and import of cultural property? There is restricted Internet access to libraries, and there are the traditional locked books. So there are negatives in the tradition.

BK: This begs the question of whether it is to society's benefit to live in this future. Do we want to make this step? Or is the status quo serving us well enough that, hey, even though technology affords us the possibility of change, let's not bother, thanks very much—we're just happy the way we are?

I'm not a technological inevitability guy. I don't believe that everything that can be built must be built. People will try to understand if they will be better off or worse off by following this path. But if that is the central issue, we are in good shape. Then we have to work with legislatures, the judiciary, and our educational institutions to make the adjustments necessary to build this future.

We have some reason to be optimistic. When I worked on an Internet publishing system called WAIS [Wide Area Information Server] in 1989, we worked with many print publishers.

The print publishers were wonderful. They started with experiments that didn't cost them very much to test the waters and then dove in. The "new media" departments in the early '90s were later melded in to become a normal part of how newspapers and periodicals worked.

You've got some false starts in the book world—the e-book stumble, which was regrettable. But it's moving along.

What the music and movie guys are doing, I can't tell you. I have not worked with many businesspeople who want to spend much time lobbying or in court.

During the '90s there was a great push to put a lot of

government materials online in the United States. We've seen a lot of that momentum erode. I wouldn't say it's partisan, but there are those who believe that broad access to information is worth the risks and there are those who prefer control. Over the years, popular support swings back and forth.

People are starting to expect that information is available online. There is a growing realization that certainly students, if not many professionals, use the Internet as their information resource of first resort—and, in fact, one study suggests, as their *only* resort in 40 percent of the cases. This estimate may even be too low. So the statement, "if it's not on the Internet, it's as if it doesn't exist," seems to be becoming true.

SF: What about being able to distinguish the best from the rest? My faculty friends comment on the ability of their students to get stuff off the Internet, but they are not able to tell right from wrong, or plausible from implausible. And, of course, every parent has a set of horror stories about materials they wish their children hadn't gotten yet.

BK: We have a couple of problems. One is that a lot of the great literature is not available online. This is a major screw-up at a societal level. Students looking around on the search engine of the moment won't find these materials. They might have to go to the restricted terminal in a public library during limited times and use a different kind of search engine, and even then things may not be available. This is no way to support a culture.

Then there's this other issue of how do you find the good stuff and separate it from the bad? I'd say that my 9-year-old has a better bull detector than I did in college. He is inundated with propaganda, and he is finding his way. Developing these skills seems to be something kids learn early these days.

I'm also quite impressed by the current search engines, where with just a couple of words, they're able to come up with a couple of answers in the top 10 out of billions of potential documents. In a flash. The technology is keeping up pretty well.

SF: Do you believe that those technologies, when applied to your universal library as opposed to the Web alone, will be satisfactory?

BK: Yes. I think we'll get more complicated than having one search engine for everything. Things will become more interesting and more complicated as the next decades roll on. But I don't think it's beyond our technological abilities to be able to pull this off.

SF: It's not just books that you will be collecting; you will likely be collecting copies of software. So your intent is not simply to publish words, but...

BK: It's everything. It has been estimated that there are 50,000 titles of packaged software. This is a doable number of items to copy onto more durable storage and provide emulated environments to be able to run them again.

We have found that preserving older packaged software is more difficult because they used copy protection for a time, but that had largely disappeared in the late 1980s. This may serve as an interesting historical note, given all the current work on copy protection, or DRM [digital rights management], for movies and music. The software industry could not get DRM to work for itself; why does the movie and music industry think the software industry can make it work for them?

If I buy Microsoft Office, I can put it on a computer, I can copy it onto another computer. There is a license key that you can put in a file. This is how the industry has grown to protect its property; it is not through digital rights management. In packaged software, it's based on law rather than technological measures.

SF: Can I ask you to go back for a moment to technology and how it works both for your bookmobile and for the Alexandria library?

BK: We're a small organization and the only way we can work well is by working with lots of other organizations. We find that we're a technology partner for others—especially in the library world where there are many experts in specific fields, but few technologists that can build and maintain petabyte machines.

Technologically, what we use for storage are Linux machines built on desktop Intel, AMD, and Via processors, with four hard drives each. Currently, those are 300-gigabyte hard drives. These are stacked up and run without modification from normal Linux. We mirror across machines and do geographic mirroring. We don't use RAID.

The Web collection grows at about 20 terabytes a month. Alexa Internet is doing most of the crawling, and we also do some with our own open source crawlers.

SF: What sort of networking do you have, and what sorts of access rates do you have?

BK: We use 500 megabits per second of bandwidth almost all the time. This is about 5 terabytes of downloads a day of mostly rich media files. The demand has been at least quadrupling each year.

The Wayback machine, which is a sort of zero-order interface to how to use this material, allows you to surf the Web as it once was. It's available on archive.org, so you can type in an URL and see past versions and surf the Web at different time periods.

That gets about 8 million hits a day, or about 100 hits per second. That's running on this Linux cluster where there's no Cisco, no Oracle, no Sun, no special anything. Everything is built out of bricks, along the Jim Gray [head of Microsoft's Bay Area Research Center] model. We do get help from people at IBM Almaden, HP Labs, Microsoft Labs—all helping to build these petabyte systems.

SF: How big is this system physically?

BK: The current active area where our machines are is about 1,000 square feet.

SF: Technologically, do you have a wish list?

BK: Yes, that we stay on track with Moore's law—that the disk guys continue at it, that the processor guys continue at it. Moore's law says that if we spend the same amount in five years we will get 10 times more of whatever it is. Probably one of the most worrisome problems—and it's actually not a technological problem—is the communications guys. The fiber engineers have been doing great work. Those guys are stripping ahead of everybody. Moore's law is nothing to them. They're awesome.

But the pricing of Internet bandwidth has not been coming down at a Moore's-law pace. I'm really worried that it's going to kill the disk-drive industry and the processor industry, because unless we move some bits around, they are going to wilt. Right now we need last-mile infrastructure, and we need a mechanism for getting to fiber somewhere closer to cost.

Again, those aren't technological problems. We just need a Moore's-law corporate mentality to spread to the communications companies—twice as much for the same money every 18 months. Many companies have done well with this approach, but if all of our industries don't stay in step, we may falter.

SF: To what extent do you view yourself as part of the open source philosophical movement?

BK: We see ourselves as absolutely part of the open source environment—though I should say that the Internet Archive, which is a tiny organization, has already spun off four companies in the last 24 months. We don't see ourselves as anticommercial in any sense. But we firmly believe open source is the best way to conduct our business. So almost all of the software we use is open source.

We are fabricating our next-generation petabyte

machine right now. Since we have paid for the metal box designs, we will make those open source (GPL, GNU General Public License) as well. The idea is to have everything from the physical hardware to the operating system open.

SF: What would you be doing if you weren't doing the library?

BK: Right now we're really settled: what we're trying to achieve is the preservation and access to all human knowledge.

One thing that I tried to do way back when was to help protect people's privacy. People in general will throw away their privacy without understanding the longer-term implications.

SF: The privacy issue is an interesting one to bring in. It can be very embarrassing to go find out what you actually said in 1985.

BK: Absolutely. We try to stay directly in touch with the thinkers in this area, so that we can make a library that has the right balances to it. I'm on the board of EFF [Electronic Frontier Foundation]. It's the ACLU of the digital world.

The Web has a lot of materials that weren't designed for eternity. If people request that they be taken out of the Wayback machine, then we do that. We try to understand where is the right balance.

SF: We haven't discussed the implications of the possibly enormous data growth that comes with video. When video catches on, suddenly more zeros may show up on some of your technology needs. The privacy implications of capturing 10 million surveillance cameras are too awful to think about.

BK: When everybody has a camera pointed at their kid's crib—do we really need to have all of that in the library? I would say we will become more selective. Right now we don't have to be selective because the technology makes it easy enough. And we're not smart enough to know exactly what it is historians want. ☺

LOVE IT, HATE IT? LET US KNOW

feedback@acmqueue.com or www.acmqueue.com/forums

© 2004 ACM 1542-7730/04/0600 \$5.00

ACM TRANSACTIONS ON INTERNET TECHNOLOGY

Transactions on Internet Technology (TOIT). This quarterly publication encompasses many disciplines in computing—including computer software engineering, middleware, database management, security, knowledge discovery and data mining, networking and distributed systems, communications, and performance and scalability—all under one roof. *TOIT* brings a sharper focus on the results and roles of the individual disciplines and the relationship among them. Extensive multi-disciplinary coverage is placed on the new application technologies, social issues, and public policies shaping Internet development. **Subscribe Today!**

PRODUCT INFORMATION

ISSN: 1533-5399

Order Code: (108)

Price: \$37 Professional Member

\$32 Student Member

\$150 Non-Member

\$13 Air Service (for residents
outside North America only)

TO PLACE AN ORDER

Contact ACM Member Services

Phone: 1.800.342.6626 (U.S. and Canada)

+1.212.626.0500 (Global)

Fax: +1.212.944.1318

(Hours: 8:30am—4:30pm, Eastern Time)

Email: acmhelp@acm.org

Mail: ACM Member Services

PO Box 11414

New York, NY 10286-1414 USA

Association for Computing Machinery
The First Society in Computing
www.acm.org

AD24

www.acm.org/pubs/periodicals/toit

AD22