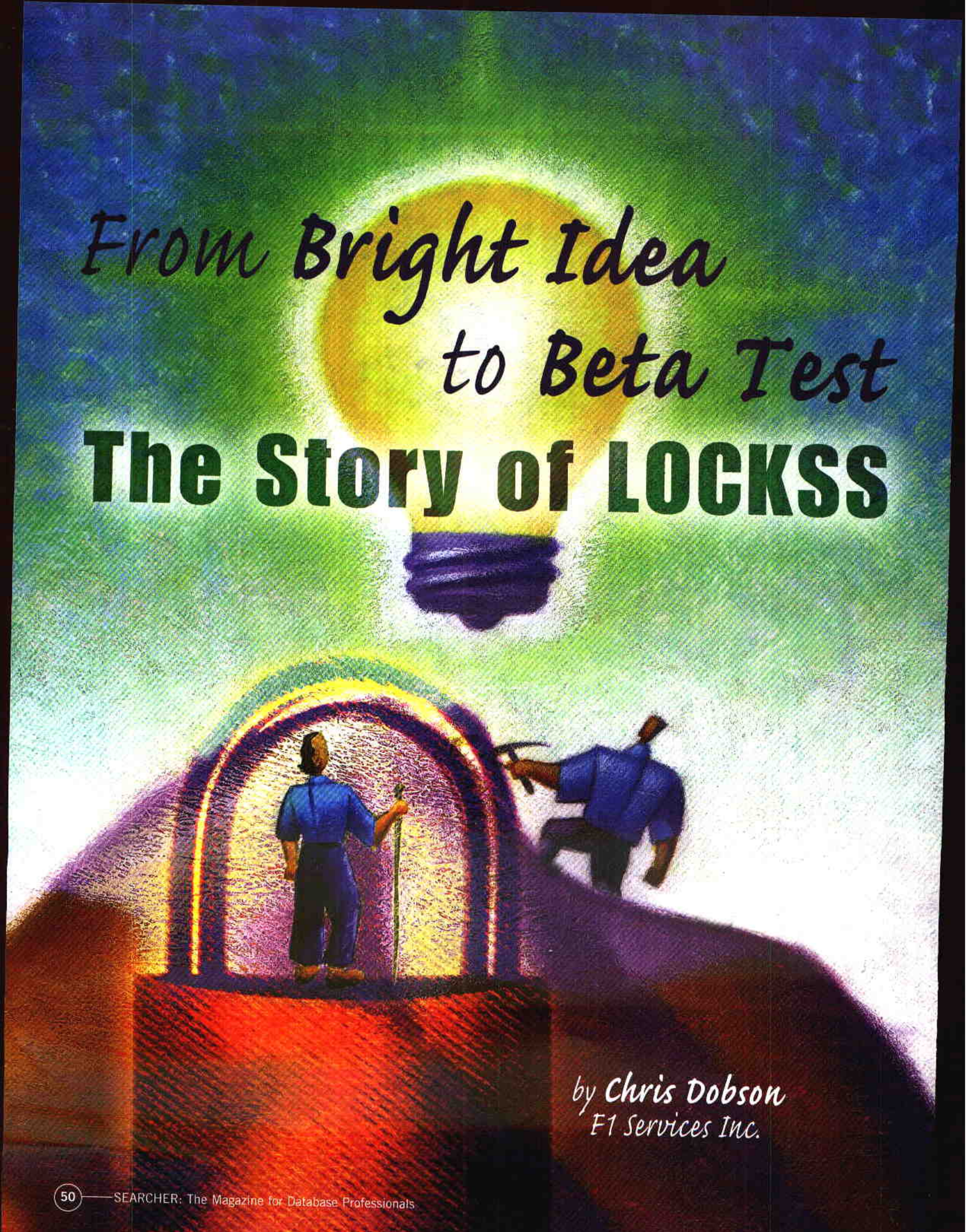
From Bright Idea to Beta Test The Story of LOCKSS



by Chris Dobson
F1 Services Inc.

Vicky Reich is a small, energetic woman who admits to bouncing when excited. But she was not excited when she met her friend, David Rosenthal, for a hike in California's Joseph Grant State Park on a beautiful Sunday in October 1998. She was irritated. As dedicated walkers know, there's nothing like a little physical exertion to focus the mind and articulate thoughts. So when David asked an innocuous question, Vicky found herself explaining at length exactly what was wrong with her world. As assistant director of Stanford's HighWire Press and a former serials librarian, Vicky was disturbed by the trends she was seeing. HighWire provides online publishing for 332 journals produced by professional societies. Working with these publishers provided Vicky with unique insight.

Publishers of journals in the medical field were gradually adding data files and other features not compatible with print to the electronic versions of their publications. Most libraries, however, lacked the funds to purchase both the print and online versions. Since the print version had always been the version of record, serials librarians were choosing print when they could only afford one version. Vicky saw problems in the future as important and sizeable portions of journals, available only online, became unavailable.

Even the casual observer can't miss the constant corporate churn of the publishing world. As a serials librarian, Vicky was more aware than most of the birth and death of periodicals and the regular trading of publications among publishers. She recognized the uncertain life of online publications.

To Vicky, the move toward online publications was cutting the librarian out of the intellectual capital loop. Libraries do not own online publications; they simply purchase access. Libraries investing thousands of dollars in content could find themselves having spent their budgets on ephemera. If a publication ceased or was sold, if a Web site was redesigned, or if the publisher's server went down, the link on the library's home page might very likely direct users to a 404 error page. Money paid for an online publication often provided access only as long as the subscription was in force and only as long as it was convenient for the publisher. Even purchasing rights "in perpetuity" was no guarantee if the publication or publisher ceased to exist. In the electronic world, it was just too easy for content to become "unpublished."

Naturally Vicky was not the only person aware of these trends. Others were concerned. Conferences and white papers have addressed the problem of archiving digital content. As a member of Stanford's Integrated Digital Library Project team, Vicky was heavily involved in the discussion of problems and options. From Vicky's perspective, however, nothing was happening. She was frustrated and ready to do something.

Intrigued by Vicky's description of the difficulty of controlling and preserving online publications, David asked how preservation worked in the print world.

Vicky described traditional paper-based system in which libraries worldwide purchase, preserve, and share content. The wide distribution and duplication of periodicals protect the preservation of intellectual property. The geographic dispersion and redundancy of the system provides protection from malicious vandals, repressive regimes, and natural disasters. In addition, the cooperative culture of libraries works to deliver specific information to whoever needs it through interlibrary loan. Those of us spoiled by the instant gratification of full text online may be frustrated by the inefficiencies and vagaries of traditional interlibrary loan. For the determined and patient researcher, however, the system works. Librarians have custody of the content and work to insure both access and preservation.

As Vicky and David reached the turning point for their hike, David said, "I can build it for you." By the time they reached the parking lot, David had outlined the technical design of a scheme that would duplicate the distributed system of print publications in an electronic environment. Vicky had a plan for raising money and she was bouncing.

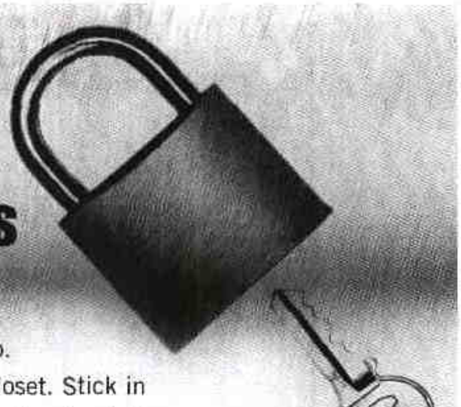
LOCKSS, Lots Of Copies Keep Stuff Safe, had been born. Of course at this early stage it wasn't LOCKSS yet. The name, and appropriate acronym, came to David later as the duo hiked in Big Basin.

Monday morning Vicky called her boss, Michael Keller, university librarian and director of Academic Information Resources at Stanford. At an afternoon meeting Vicky introduced David to Michael and vouched for David's expertise. David Rosenthal is currently a distinguished engineer at Sun Microsystems. In 1998, he was involved with a startup company. After working at Sun for 8 years, he had left to work first with one startup and then a second. Although both startups succeeded, the grueling pace was tiring and David was starting to think about a change. Vicky received Michael's support for the project, which included allowing her to split her time between HighWire and LOCKSS. Now all she had to do was raise money.

Raising Consciousness and Money

As Vicky tells the story of LOCKSS, the funding and support from both libraries and publishers are the product of happy circumstance. In reality, they are the result of dozens of presentations, workshops, and assiduous networking over several years. Vicky was an expert on serials pricing, digital libraries, and the ways people interact with libraries well before she took on the task of preserving electronic journals. Her intelligence, determination, hard work, and accomplishments at HighWire yielded a receptive audience. Early on she made contact with Michael Lesk, a colleague who had just accepted a position as division director for Information and Intelligent Systems at the National Science Foundation.

How LOCKSS Works



Take a computer a generation past its prime, not a Sun workstation, a Dell will do. Hook it up to the Internet and put it in a closet. Stick in the LOCKSS CD-ROM and boot it up. Close the closet door.

It's almost that easy. The LOCKSS design facilitates control by librarians, not the information services department. The design team has also considered the budgetary constraints of libraries. The goal is an effective system that will run on hardware scrounged by resourceful librarians with almost no administrative overhead. It runs on LINUX, an open source (free) operating system, and the LOCKSS software itself is freely available. The design team hopes to foster a community of librarians and library systems administrators that just won't be able to keep their hands off LOCKSS — tweaking it, improving it, and collaborating with each other to keep it up-to-date and effective long after the original team has moved on.

The LOCKSS design looks a bit like the Internet itself. Little nodes (the computers in the closets) are spread all across the world, connected to each other, with messages constantly moving from one to another. Each LOCKSS computer downloads everything contained in an issue of a subscribed journal from the publisher's site. Every few weeks, the LOCKSS computer compares the issue it has with the same issue at a random selection of other LOCKSS computers. If any differences are found, the damage is repaired by downloading a fresh copy from the publisher or one of the other LOCKSS computers. A hardware failure or a security breach could cause differences between LOCKSS caches. Regardless of the cause, the damaged data is restored. The LOCKSS computer stays in the closet, quietly collecting issues of a publication until disaster, a corporate merger, or a hacker strikes the publisher's Web site. Then the library taps into the LOCKSS computer to pull up the publication without dropping an issue.

Although the basic design is firmly in place, lots of work remain. The beta test proved that the software can effectively transfer publication issues to the distributed LOCKSS caches and recover data following a failure. The remaining development tasks include the following:

- Increasing the speed with which LOCKSS gathers data from publisher sites
- Making the software flexible enough to run on a number of different platforms
- Providing a better user interface for setting up the LOCKSS computers
- Devising ways to connect the LOCKSS computers to the institutional systems for those times when publisher sites go down.

Additional details on the inner workings and requirements for beta test sites are available at http://lockss.stanford.edu. If you'd like to be involved in the Phase 2 beta test of this exciting project, contact Vicky Reich at vreich@stanford.edu. ❖

Through the National Science Foundation's (NSF) Small Grant for Exploratory Research (SGER), Michael could provide the initial $50,000 funding to take LOCKSS from a bright idea to the alpha test stage. When David rejoined Sun Microsystems in 1999, he continued to work on LOCKSS with the support of Sun.

With initial funding secured, Vicky hit the seminar trail. In June 2000 she and David presented a paper at the USENIX Conference and in December at Preservation 2000 in York, England. Initial reaction to LOCKSS was mixed. The tortoise logo suggested by Vicky may have reflected not only the deliberate slowness of the system architecture, but her realization that getting LOCKSS up to speed would take some time. Vicky and David had spent the fall and winter of 1998-1999 working on details, using time stolen from their primary jobs. Not until spring 1999 did they have the opportunity to complete the paperwork for the NSF grant.

Definition and Development

The original scheme that David had devised mirrored the existing system for print collections in its complete decentralization. Numerous libraries would cache copies of their electronic subscriptions for use only when the publications became unavailable from the publisher. The key innovation was development of a protocol that would poll the dispersed caches to ensure that none had been corrupted and to correct any errors found. LOCKSS would actually go one step beyond the system for print archives by automatically restoring content in the event of a failure. While a fire might devastate a collection, once a computer was connected to the Internet and loaded with the LOCKSS software, the library's electronic collection could be completely restored, for free, after just a few weeks.

The technical and administrative challenges were tremendous. The basic system software had to be developed and had to run on computers libraries could afford. How to protect publisher rights? What content to cache? How the LOCKSS cache would interact with the library's systems when the publisher site became unavailable? How to protect the system from hackers? All these issues and others had to be addressed. In just a few months, the team was ready to begin serious testing of the concept and the software. The alpha test in 2000 used content from AAAS Science Online and involved libraries at Stanford, the University of California-Berkeley, Los Alamos National Laboratories, the University of Tennessee, Harvard, and Columbia.

Even with the alpha test underway, many in the library community regarded LOCKSS as a technology project rather than a real archiving solution. The features that make it most unique undoubtedly contributed to the skepticism.

- Unlike most archiving projects, LOCKSS does not have a centralized repository. The distributed nature of the archive is its key principle.
- The administrative structure is decidedly light weight. The goal for 2004 is a production system and support organization that can be largely self-managed by participating libraries. All the LOCKSS software is open source and available to anyone to use or modify.
- LOCKSS appears to provide leading-edge technology but operates on old machines with minimal administration. The idea of slipping a CD-ROM loaded with free software into an outdated computer stuck in a closet and walking away until the publisher's site goes down seems farfetched. [See the "How LOCKSS Works" sidebar at left.] A key concern of the LOCKSS team is making the system affordable both in terms of hardware and administrative time.

The alpha test revealed an encouraging robustness. The system survived a fire at Los Alamos, network problems at Stanford, relocation of the computers at Berkeley, and flaky hardware at Columbia. The time had arrived to seek serious money. Again, LOCKSS benefited from fortuitous timing. Michael Keller learned that Don Waters at the Mellon Foundation was actively seeking electronic journal archiving projects. Knowing that funding is available, however, is only the first step. Applying for grants forced the team to carefully refine and articulate the goals of LOCKSS and the operational plan. New priorities were set and the beta test plan matured. As they considered grant proposals, Vicky continued talking and writing about LOCKSS. She has made more than a dozen presentations on LOCKSS in 2 years, alerting librarians to dangers they hadn't yet realized, describing the unique LOCKSS system, and soliciting participants for the beta test.

Because the beta test needed to be much larger than the alpha test and international in scope, Vicky also went back to her network. Through friends, acquaintances, the Council on Library and Information Resources, and the Digital Library Foundation, she found 50 libraries, at least one on every continent except Antarctica, willing to participate. Although the beta test uses simulated journals, Vicky has been talking with publishers. BioMed Central, Blackwell Publishers, American Physical Society, American Chemical Society, and the Nature Publishing Group have expressed interest in participating in the project. As the beta test has progressed, even the most skeptical have come to appreciate the soundness of both the technology and the concept.

As the first phase of the beta test drew to a close in the summer of 2002, the LOCKSS team celebrated successes and identified targets for future development. The beta test proved that the fundamental design was sound. The system collected, preserved, and repaired content. The test also convinced the designers that the system would continue to operate efficiently, even with significant increases in the number of sites and the number of publications. Now the work of adding efficiencies to the system and developing interfaces between the LOCKSS computers and the host libraries' systems begins. An additional grant from the Mellon Foundation will enable the LOCKSS team to embark on Phase 2 of the beta test and take the system into the real world. The NSF has also provided additional funding to continue the research on the underlying technology. If you are interested in participating, you can find detailed information at http://lockss.stanford.edu or you can contact Vicky at vreich@stanford.edu.

Beyond the Beta Test

Although the technical development of LOCKSS is proceeding with reasonable speed, Vicky has not slowed down or slacked off. She's busy campaigning for additional beta test participants. She is also talking about LOCKSS to nonmedical publishers. Some smaller publishers are reluctant to participate in some of the other archiving projects because they don't have the capability of packaging the content to conform to the system specifications. Because LOCKSS gathers data directly from publisher servers, the barriers to participation are minimal. Vicky plans to engage publishers outside the scientific/technical/medical community. She sees a critical need for LOCKSS in the humanities, where many publishers are small and library funds tight. She has already begun talking with the New York Public Library at Lincoln Center about preserving electronic performing arts collections. Vicky also wants to involve collection development librarians in identifying and promulgating best practices for selecting and building local digital collections. Although the problems with electronic journals that disturbed Vicky in 1998 remain, she has confidence that LOCKSS will be a solution. ◆