JUDE BUFFUM

by ARIEL BLEICHER

A MEMORY of WEBS PAST

EVERY WEEKDAY AT 5:00 A.M., A NONDESCRIPT GRAY VAN rolls down the underground service road beneath the French National Library, in Paris, and arrives at a svelte glass skyscraper soaring above the bustling Seine River. Here, at the Tower of the Times, the van delivers a tiny but astoundingly rich snapshot of life in this country that takes its cultural heritage very seriously.

The van has been stuffed willy-nilly with two copies each of some 3000 periodicals printed recently in France that are being sent to the library for preservation. One morning last November, the haul includes the dailies *Le Monde* and *L'Humanité*, of course, and also the union newspaper *Le Travailleur*. Among the other lexical artifacts dutifully funneled from the van up into the tower are a booklet of classified advertisements, a concert flyer, several religious pamphlets, *Busty Beauties* magazine, and a community newsletter from Bonnes (population 330) announcing a town raffle for three hams, six bottles of wine, and a yogurt-making machine.

The Web is a rollicking, revealing record of life in the 21st century. But preserving it for future historians is a monumental technical challenge

"We have a lot of so-called crap, and we're happy about that," says Gildas Illien, an archivist at the library. His colleagues in other countries might turn up their noses at hard-core porn, advertisements, and obscure newsletters, but not Illien. "In a hundred years, what's totally irrelevant or dirty today will end up becoming of extreme interest to historians," he declares.

THE TOWER OF THE TIMES, where Illien works, is one of four spires, each composed of two perpendicular wings resembling the pages of an open book, that make up France's newly modernized national library. The archivists here aren't after just printed material; they're preserving the electronic, too. In fact, it's Illien's daunting task to archive French Web sites—all of them, in all their evanescent, constantly changing, and multimedia splendor.

Since the ancient Sumerians compiled the first collections of inscribed clay tablets, many peoples have attempted to preserve documents, ephemera, and even the flotsam of their political, economic, and social tides. But perhaps no nation today tackles this endeavor as thoroughly as France, one of the few countries in which archivists have the legal right to copy and save virtual documents without fear of a copyright suit. Five centuries ago, King Francis I ordered book publishers to donate copies of their work to posterity. That legal deposit law, as it is known, has expanded over the years to include maps, music scores, periodicals, photographs, sound recordings, posters, motion pictures, television broadcasts, computer software, and finally, in 2006, the World Wide Web.

French archivists are still grappling with that most recent mandate. The Web, of course, is unlike any other publishing platform—not simply because it is amorphous and immeasurably large but because its "documents" are boundless. Nowadays, an "online publication" is barely recognizable as a publication in any traditional sense; it exists in a perpetual state of being updated, and it cannot be considered complete in the absence of everything else it's hyperlinked to. Unlike books and newspapers, which have discernible titles, authors, beginnings, and ends, the Internet is utterly nonstandardized.

The task of preserving what's put online has proved, to no one's surprise, monumental. And it's only getting more so as the Internet expands, as Web sites become more dynamic, and as concern grows over online privacy. Increasingly, much of what people put online is being diffused across social networks and distributed through personalized apps on smartphones and tablet computers. The classic Web site, it seems, is already starting to slide toward obsolescence. "I'm convinced the Web as we know it will be gone in a few years' time," Illien says. "What we're doing in this library is trying to capture a trace of it." But to do even that is requiring engineers to build a new, more sophisticated generation of software robots, known as crawlers, to trawl the Web's vast and varied content.

ILLIEN SEES HIMSELF AS A STEWARD of an ancient tradition; he believes he is helping pioneer a revolution in the way society documents what it does and how it thinks. He points out that since the end of the 19th century, the French National Library has been storing sales catalogs from big department stores, including the famous Galeries Lafayette. "Today," he says, "this exceptional collection...is the best record we have of how people dressed back then and who was buying what." One day, he insists, the archives of eBay will be just as valuable. Capturing them, however, is a task that's very different from anything archivists have ever done.

A CULTURAL REPOSITORY: Every day, thousands of books, periodicals, brochures, and street flyers pass through the sorting rooms [above] in the basement of France's national library [left]. Eight stories above, a team of Web archivists hunt down digital documents and archive them in PetaBox storage servers [top] designed by San Francisco's Internet Archive.
PHOTOS: BIBLIOTHÈQUE NATIONALE DE FRANCE

The Web is regularly accessed and modified by as many as 2 billion people, in every country on Earth. It's a wild bazaar of scripting languages, file formats, media players, search interfaces, hidden databases, pay walls, pop-up advertisements, untraceable comments, public broadcasts, private conversations, and applications that can be navigated in an infinite number of ways. Finding and capturing even a substantial portion of it all would require development teams and computing resources as large as, or probably larger than, Google's.

But Google, aside from saving previously indexed pages for caching, has mostly abandoned the Webs of the past—the complete set of Web pages as they existed a month, six months, a year ago, and so on, back to a site's origins. Thus the job of preserving them has fallen to nonprofit foundations and small, overworked teams of engineers and curators at national libraries. Illien, for example, manages a group of nine.

For a digital archive, the French National Library's collection of Web data is surprisingly small—just 200 terabytes stored on hard disks and magnetic tape in the library's data center. It includes copies of French Web pages dating back to 1996. Illien's team completed its first harvest of the entire French domain (.fr) just last summer. Other national libraries, such as Iceland's, have been downloading their national domains periodically since the early 2000s.

Part of the difficulty in fetching the contents of the Web is that no one really knows how much is out there to be fetched. Brewster Kahle, a U.S. computer engineer who in the late 1980s invented the Wide Area Information Servers, a pre-Web pub-

SPECTRUM.IEEE.ORG

lishing system, paid a visit to AltaVista's offices in Palo Alto, Calif., in 1995. He was shocked to see that the then-popular search engine had indexed 16 million Web pages "on a set of machines that were the size of two large Coke machines," he recalls. "You could actually wrap your arms around the Web."

The apparent compactness of the Web inspired Kahle to found, in San Francisco in 1996, the nonprofit Internet Archive. Wary of infringing on copyrights, AltaVista made sure to delete old pages in its cache. But the Internet Archive, emboldened by its status as a trustworthy nonprofit, was willing to be brazen. "We have an opportunity to one-up the Greeks," Kahle says, referring to the ancient philosophers who collected hundreds of thousands of papyrus scrolls in the great Library of Alexandria. The invention of the Internet, he argues, has made it possible to create an archive of human knowledge that anyone can access from anywhere on the planet. And Kahle, for one, wasn't going to let a bunch of lawyers talk him out of it.

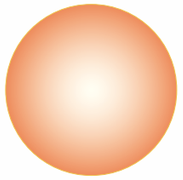
By March 1997, he had compiled what was arguably the first true time capsule of the global Web. In fact, a substantial portion of the French National Library's electronic archive was simply bought from Kahle's Internet Archive. One of the archive's major successes has been its online access interface, called the Wayback Machine, which lets anyone who knows the address of a Web site see archived versions of its pages. Today the Internet Archive stores more than two petabytes of Web data in a portable Sun Microsystems (now Oracle America) data center built into a shipping container. Back in 1997, Kahle had captured nearly 2 terabytes, which he calculated was about a tenth the amount of text stored in the entire U.S. Library of Congress. It was a substantial collection of the Web of the time, but it wasn't nearly everything.

Kahle knew there were still hundreds of thousands of sites and perhaps millions of "hidden" documents, images, and audio clips that his crawler program missed. It couldn't access password-protected sites, for example, or isolated pages with just a few if any hyperlinks, such as outdated product postings on eBay. More troubling, it couldn't probe "form-fronted" databases, which require typing keywords in search boxes to call up information (such databases include those at the National Climate Data Center in the United States and the British Census). Still, Kahle believed that with the right tools and enough human curators to guide the crawlers, it was possible to get almost all online data. The Web may have been big, but ultimately it was manageable.

That is no longer the case. The part of the Web indexed by search engines such as Google has ballooned from some 50 million unique URLs in 1997 to about 3 trillion today, according to the latest update last November by Majestic SEO, a search optimization service. A URL, or uniform resource locator, designates a single document, such as a JPEG image or an HTML text file. Those files, however, are just a tiny piece of the Internet. By some estimates, the total "surface" Web visible to crawlers is six times the size of the indexed Web, and the "deep" Web of hidden pages and databases is some 500 times larger still.

INDEXED UNIQUE URLS

1997
50 million



2010
3 trillion

COUNTING URLS, THOUGH, has become a fairly pointless exercise. For instance, it's possible and increasingly common that a single site is capable of generating vast numbers of unique URLs, all pointing to the same content: advertisements or pornography, typically. Though engineers have devised tricks for steering crawlers away from such spam clusters, even Google's crawlers still from time to time capture billions of unique URLs redirecting to the same place.

"In reality, the Web is infinite in all the wrong ways," laments Julien Masanès, who introduced Web archiving at the French National Library in 2002 and managed the collection until 2004, when he left to start what is now the nonprofit Internet Memory Foundation, headquartered in Amsterdam and Paris.

Realizing the Web's anarchic nature, its early archivists quickly gave up trying to dig up documents from every nook and cranny and instead focused on making quality copies of the pages they knew they could find. After all, they were building historical archives, not indexes of live sites, and it was imperative that their crawlers retrieve complete and perfect replicas. The trick was having a program that could fetch everything—not just text, which was what most search crawlers prioritized, but also images, graphics, and video.

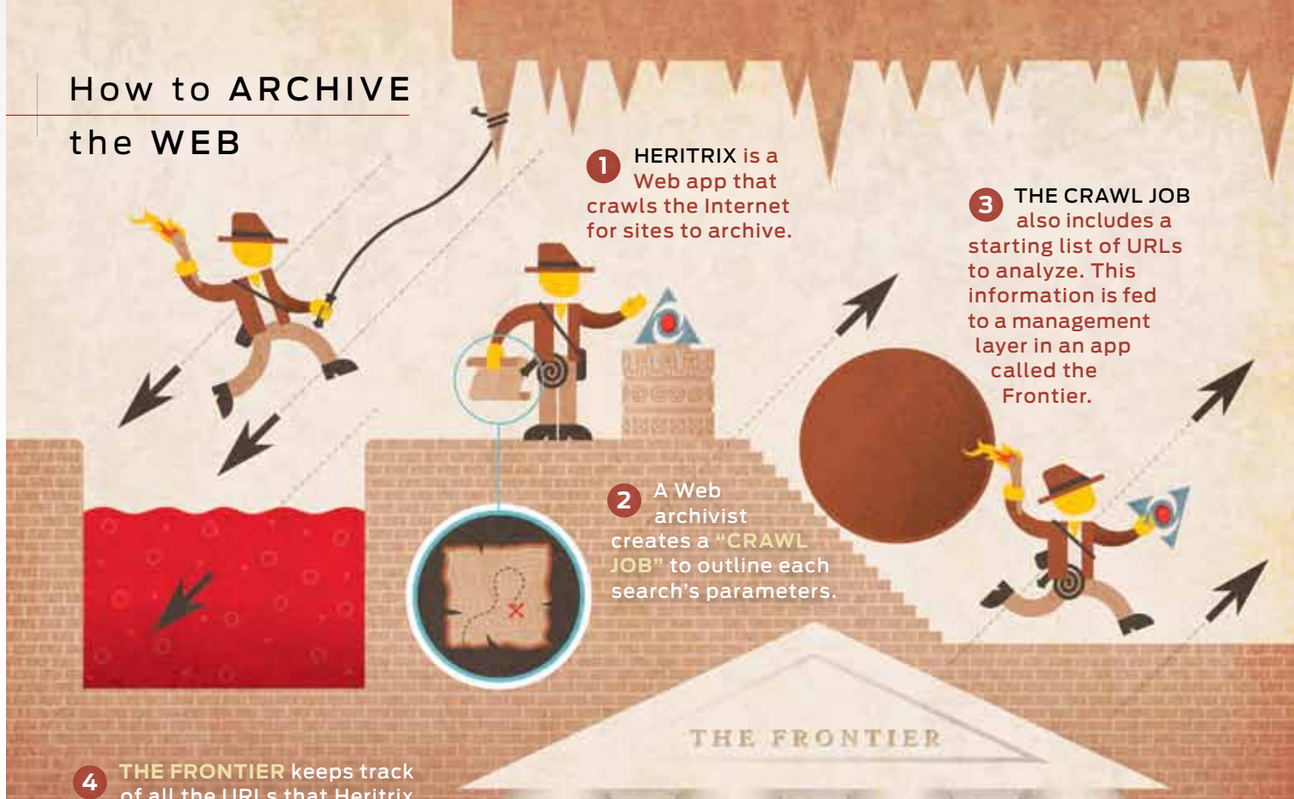
When Kahle first started saving copies of Web documents in the late 1990s, he was trawling the Web with a crawler he helped develop for Alexa Internet, a search company he founded in 1996, the same year he established the Internet Archive. But three years later, he sold Alexa to Amazon.com, along with the rights to its software. No big loss, he figured. Alexa would still donate its twice-yearly global Web crawls to the Internet Archive, and in the meantime, Kahle and his engineers would build a crawler that was open source, meaning that anyone wanting to use or modify the software could download it for free. "Companies come and go," Kahle says, and because the goal is to build an archive of the Web that would last indefinitely, "the idea of depending on corporately controlled software is not a long-term strategy."

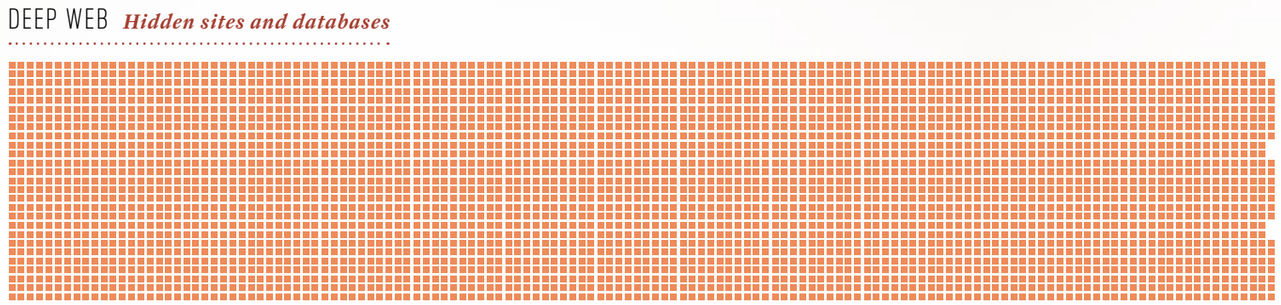
So Kahle hired a young Internet software developer and self-described "steward of open source" named Gordon Mohr to take charge of coding the crawler that would ensure the world's digital inheritance. Mohr had few good models to work from. "In the earliest days of search crawlers, an awful lot of them immediately reduced a site to plain text," Mohr notes, explaining that the "index quality" crawlers weren't made to preserve a site's "original appearance and functionality." But in January 2004, he released the first public version of his "archival quality" crawler and named it Heritrix, an archaic synonym for "heiress."

BEFORE HERITRIX, FEW LIBRARIES in the consortium had developed the technology to do any real archiving of their own. The problem was mainly a lack of resources. "We're too small, we're not smart enough, and we're terribly French," Illien explains, only half joking. Most libraries, including the French

JUDE BUFFUM

How to ARCHIVE the WEB





National Library, have eagerly adopted Heritrix. But Heritrix is far from a turnkey solution. Configuring it and guiding it through cyberspace require nontrivial engineering mastery and on-the-job innovation. It's not at all like transferring your iTunes library from one computer to another.

"It's more like playing, eh, you know the game *Tetris*?" says Annick Lorthios, who works with Illien at the French National Library. Part of her responsibilities is monitoring Heritrix as it harvests things from cyberspace, which in order to crawl through all things French (stuff on the .fr domain, mostly) takes about two months. It sounds fun, except in this game, if you lose, either your system crashes or you end up storing a lot of junk you don't want, like 5 billion copies of a credit card ad.

Getting Heritrix to fetch the things you do want, explains Sara Aubry, who leads the two programmers on Illien's crew, is about making rules—"setting parameters," she calls it. For example, she can order the crawler to "Stay inside this domain" or "Don't even think of downloading more than 80 terabytes of data" or "For heaven's sake, stay away from URLs that look like *those!*" Then she gives the robot a list of places to go—about 1.6 million URLs purchased from the French domain registries, for example—and sends it on its way.

Heritrix enters cyberspace through this seed list of URLs, which, like street addresses, tell the robot where documents reside. In an ideal world, the crawling process happens like this: Heritrix follows each URL to the server that's storing the document and asks the server, "Do you have this document?" The server responds, in effect, "Yes. Here it is." Heritrix downloads the document, then scans it for more URLs. If those URLs lead to real things and Heritrix hasn't seen them before, they go in a "things to be downloaded" queue—one for each server the robot visits.

As Heritrix downloads the documents in its queues, it parses them for still more URLs, zipping from server to server, diffusing across the Web. Simultaneously, the crawler adds a few notes about where it found the documents and when, then stuffs them in big "suitcase" files, which are themselves piled up, compressed, and stored on disks.

Rarely, though, do things go so smoothly. The Web is a nasty place for a crawler, full of "crawler traps." A Web site with a calendar, for example, can unintentionally stall a crawler and keep it from fetching useful things. If each page of the calendar generates a link to the next day or the next month, it will create new URLs for every date until eternity, and "stupid Heritrix," as the Internet Memory Foundation's Masanès says, will ask for them all, one by one.

Sites that intentionally spam, known as spam clusters, are much more sophisticated. They involve heavily cross-linked networks of content that's often stolen or copied from other sites. The pages of a spam cluster all cross-reference one another, creating the illusion that a lot of people are linking to a site. The upshot for the spammer is that if Google's crawlers fall into this trap and index the site, its page rank improves dramatically, which makes Web surfers more likely to find it and click on it.

Such crawler traps are an archivist's nightmare. Let your crawler fall into a few and your archive is quickly spammed with billions of worthless files. Let your crawler fall into too many and the computing power needed to deal with such a large pool of URLs can overwhelm your servers and crash your system. The traps are why Illien's coworker Lorthios thinks of monitoring Heritrix as like playing a round of *Tetris*: Let too many blocks of the wrong shape stack up and your screen fills before you can win any rows. Game over.

For big crawls, it's easy to miss noticing that your crawler is gathering spam; the software downloads so many URLs so quickly that you may simply overlook a chain of suspiciously similar URLs in one of thousands of queues. Web archiving engineers can code special spam filters for Heritrix. Yet spammers are always inventing new tricks, and no mathematical method can warn Heritrix about them all.

The variability of Web formats has become a big problem for Heritrix, not just for avoiding traps but for capturing content. When Mohr designed the crawler's original architecture in 2003, the Web consisted mostly of pages of HTML text. "A Web page was just a file and everything was in the file," says Jérôme Thièvre, a software engineer at the French National Audiovisual Institute, in Paris, which archives French television and radio, including Web broadcasts.

Heritrix had no problem finding documents in a file; that was what it was built to do. But as the Web evolved, it grew into "a kind of jungle in terms of technology," Thièvre says, and archivists are particularly worried about being able to capture its newest design fad: rich media.

Rich media is pretty much anything that moves when you interact with it. It can be a Flash animation, or a streaming video clip, or an image rotator that changes when you mouse over it. The pages describing this kind of media are coded in such a way that your actions within a browser—clicking a Play icon on a video player, for example, or typing an address into Google Maps—generate or trigger visible content. But Heritrix, because it looks for ordinary HTML files, fails to recognize the more dynamic components of these pages. So when Heritrix crawls sites heavy in rich media, it can miss as much as 40 percent of their content.

A few developers, particularly those at the Internet Archive and the Internet Memory Foundation, are experimenting with ways to get around this problem and patch the holes in their archives. They are building supplemental crawlers that act more like browsers, for example, or configuring Heritrix to work collaboratively with other downloading programs, and they have had some success. But most archivists lack the servers and funds to develop new tools and are simply doing the best they can with the ones they have.

"Right now, we're 100 percent ready to archive the way the Web was 10 years ago," Aubry says. "You know, plain HTML pages, nothing's moving around, not a lot of video—just images and text."

EARLY ARCHIVISTS NEVER ANTICIPATED that their biggest obstacle to building a comprehensive archive of the most accessible knowledge base in history would be providing access.

In most countries, including the United States, legal deposit laws don't apply to the Web. Copyrights, on the other hand, do. So in the strictest interpretation of copyright law, it's illegal for anyone, even a national library, to make and share copies of an electronic document, whether it's a music file or an online news brief. "Plenty of newspapers are earning money from charging readers to access their archives," explains a lawyer for the Library of Congress. "They could lose that money if we provide the content."

"That said," interjects Abigail Grotke, who leads the library's Web-archiving team, "it really crimps our Web-archiving style." In respect of U.S. copyright law, Grotke's team archives only government Web sites and several thousand select sites whose publishers have sent written consent.

The risk of inviting copyright lawsuits has driven other institutions to create "dark archives"—copies of their complete national domain that no one can see—with the hope that eventually the law will change. "I worked for six years putting things in the box without giving access to all that data," says Aubry, who was hired at the French National Library in 2002. "We used to call it the 'black box.' The first day we opened our collection [in 2008] was a happy day."

Still, the French National Library's Web archive isn't accessible to everyone. In fact, the archive is open only to researchers and browsable only through the computers in the library's reading rooms. Illien worries that if he tries to make the archive publicly accessible through the Web, France's personal data protection agency (Commission Nationale de l'Informatique et des Libertés), which hasn't yet legislated on Web archives, will step in and limit access even further—something that's already happened in Denmark and Norway. The agency strives to help French citizens retain control over their own information—to protect the teenager, say, who naively published pictures of herself on her Web site and doesn't want future employers to see them.

Most Web archives are similarly restricted, which frustrates idealists like Kahle. The Wayback Machine has existed online without controversy since 2000, he points out. And although the Internet Archive will remove a site from the Machine at the owner's request, few people have asked. "If we don't want to lose what it is we've built as humans—this enormous effort of putting knowledge on the Internet—we've got to go and not only capture it but make it accessible again," he insists. When he helped establish the International Internet Preservation Consortium eight years ago, he had hoped the national libraries would lead the charge. "Frankly, they've failed," he says. Indeed, while the Wayback Machine receives an estimated 400 000 unique visits a day, the Web archive at the French National Library gets just 80 users a month.

Though Illien would like to see his archive go online someday, he doesn't see the point in rushing things. "Users will come when the Web is dead," he declares as he waves me through a security desk and into the library's reading room. It is dimly lit, with a lofty, arching ceiling, and creepily quiet.

Crouching over a computer terminal, he shows me an early selection from one of his favorite collections, which consists of prototypical French weblogs. Illien loves how this archive distills the essence of the Web's larger evolution. "Early on, only computer geeks could write a blog," he whispers, grinning at the screen. "So the first stories of the Web were stories of the ordinary life of nerds. Then Web sites became more accessible, and you get love stories, travel diaries, people writing about their lives from the wildest parts of society."

He selects the blog, from 1997, of a computer science student. It's titled "MöngôlO's Diary (Almost)." The first entry begins "*Une de mes grandes phobies est de ne pas être compris*": "One of my greatest fears is not to be understood." You can almost picture MöngôlO, hunched like Illien over a boxy gray monitor, trusting that the Web would free him from oblivion and misunderstanding. By 2001, his blog had vanished off-line.

But it's not gone for good. It still exists (almost), stored safely as electronic bits inside a whirring machine room in a library in Paris. ❑

JOIN THE DISCUSSION at http://spectrum.ieee.org/webarchive0311.