Electricityng knowledge

An effort to scan the world’s books is under way — creating a growing, globally accessible virtual library that had its start in Pittsburgh.

In 1995, Carnegie Mellon University’s Professor Raj Reddy organized a meeting in Shadyside of the world’s foremost digital thought leaders to discuss the feasibility of electronic libraries. The idea of very large Internet libraries had been gestating in Reddy’s mind for about 15 years, but it was not until then that desktop computers, easy Internet access, Web browsers, image scanning and optical character recognition became sufficiently widespread to allow his vision to become a reality.

Known around the world for his vision in artificial intelligence and human-computer interaction, Reddy founded the Universal Digital Library (UDL) and set about enlisting the people to make it work. His first recruit was Jaime Carbonell, a CMU linguist, computer scientist and mechanical translation expert. Several years later, in 1998, Reddy recruited Michael Shamos, physicist, computer scientist, technology manager and attorney, to deal with the manifold legal and executive tasks associated with the undertaking. Shortly thereafter, CMU’s newly arrived dean of libraries, Gloriana St. Clair, introduced herself to Reddy. Having come from Penn State, where she had inspired and directed the migration of the statewide library system’s science and technology resources from campus-specific paper subscriptions to system-wide digital licenses, St. Clair had no idea about Reddy’s visionary venture. “When I met Dr. Reddy, I sat down and said, ‘You know the future of libraries is digital.’” She paused and smiled, adding, “It was a coals-to-Newcastle moment. So he asked me to become involved in the Universal Digital Library, and after a while I became one of the directors.”

Today, Reddy, Carbonell, Shamos and St. Clair serve as co-directors of the UDL, an international collaborative dedicated to making all human knowledge available to all humankind. Broadly speaking, Reddy makes the international deals, St. Clair manages library resources, Carbonell makes the technology work, and Shamos handles intellectual property issues. Another director on leave at this time is Robert Thiibideau, chief scientist at Seagate, Pittsburgh.

As a practical matter, the UDL’s self-directed charge is to locate, scan, index and post all published works of humankind, both ancient and modern, accessible and obscure, free for reading on the World Wide Web. The directors estimate that of approximately 100 million works...
SURFING THE STACKS

Physical access to library buildings may become unnecessary when scans of book pages are available online.
ever produced, 50 million are likely to be found.

The project's ambitious proportions are tempered by the sober recognition among its leaders that achieving their objective within their lifetimes is extremely unlikely. They reckon progress not in years and decades, but in generations and centuries. In late 2007 the UDL celebrated the successful completion of its proof-of-concept experiment, called The Million Book Project, in which, as the name suggests, they successfully scanned the first million of the many millions of books that will someday populate the library.

Organizationaliy, the UDL comprises four entities: Carnegie Mellon University and the governments of India, China and Egypt. With $3.5 million in grant funding from the National Science Foundation, the project has purchased equipment for 28 scanning centers in China and 21 in India. The U.S. contribution to the enterprise is small compared with the full cost of operations.

"India and China have been putting in about $10 million in kind a year," Reddy said. "They have five or six hundred people for whom they pay salaries, provide space and pay for utilities. They have people finding the books, bringing them in and taking them back. Giving them a computer and scanner is less than 5 percent of the cost. So the leverage we're getting is enormous."

If the UDL sounds like an unqualified success today, it wasn't always that way. According to Shamos, the project started out more modestly than its recent success would suggest. "One of the first things we did when I came to the project in 1998 was the Thousand Book Project, in which we found the right scanning equipment, the right resolution to do the scanning, [decided] whether to cut the books up or not, and [chose] the best formats to display them in."

By 2000, with those questions adequately answered, the next logical step was to scale up to The Million Book Project. A multi-year quest for funding ensued. Reddy described the prevailing attitude of funding sources at the time: "When the proposal was sent out for reviews, half the people said, 'This is great,' and half the people said, 'This is a pipe dream that will never happen.' So after many years of trying, in 2002 we got a $3.5 million grant from the National Science Foundation, which we used for equipment that we gave to China and India for scanners and computers."

For Carbonell, the UDL is an answer to the historic and continuing problem of the loss of human knowledge, both intentional and inadvertent. "A long time ago the world's greatest tragedy was the burning of the Library of Alexandria," Carbonell said. "That was a single point of failure for a huge treasure trove of human knowledge. With the Universal Digital Library, that will never happen again."

According to the Roman historian Plutarch, in 48 B.C. the Roman Emperor Julius Caesar set fire to his ships at Alexandria in an attempt to evade their capture by his enemy, Egypt's King Ptolemy III. As fortune had it, Caesar's ship-torching tactic precipitated a calamity of monumental proportions. Flames from the burning vessels ignited the dock, which in turn set the world's greatest repository of knowledge, the Royal Library of Alexandria, ablaze. The Roman philosopher Seneca reported 40,000 books having been burned in the fire. In a testament to the tenacity of the human thirst for knowledge, the library managed to rebuild and flourish for the next seven centuries, surviving at least three subsequent sackings and burnings, but finally succumbing to the ravages of combustion during an attack in A.D. 640. As a consequence of its repeated misfortune, the ancient library at Alexandria has become emblematic of the destruction of knowledge that has occurred periodically for millennia around the world.

Today, UDL's Egyptian partner, the New Biblioteca Alexandrina (BA), responds to its own historic legacy as well as to the modern-day mission of its international partners as being "dedicated to recapturing the spirit of openness and scholarship of the original Biblioteca Alexandrina." Completed in 2002, the architecturally inspirational library is a monument to the spirit of human knowledge.

Time travel at the library

A search for Caesar's enemy, Ptolemy, on the UDL Web site returned no references to the king. Serendipitously, however, the index returned a 1562 translation of the "Geographia," by Claudius Ptolemy of Alexandria (probably not related to the regent). This Ptolemy, who was the western world's first geographer and astronomer, lived in Hellenistic Egypt between 83 and 161 A.D. It would have been during the ancient library's second incarnation that Claudius Ptolemy might have meandered among the stone pillars and wooden ladders of the stately edifice, perusing the scrolls of the library's collection as he researched his great works, "Geographia" the first atlas of the earth, and "Almagest" the first atlas of the heavens.

As indicated by its title page, the Latin translation of the 1,500-year-old Greek
original "Geographia" by one Bilić Perckheyer, had been emended by the mathematician Joseph Moletio and was published in Venice in 1562 by the publishing house of Vincent Valgris. The book's title page is stamped with Saint Francis of Assisi's chosen emblem, the Tau Cross, which is circumscribed with the name of the Capuchin monastery of Verona, Italy. Verona was presumably the book's long-time residence before traversing the Mediterranean to return to its birthplace, Alexandria, Egypt, where it became part of the Biblioteca Alexandrina and was subsequently scanned, digitized, indexed and posted to the UDL. The online search of www.ualib.org took less than a minute. Ten minutes later, the 692-page Acrobat file of the book had made its way from Alexandria to my desktop in Pittsburgh.

While the 1562 version of the book is in the form of a codex, like the bound books we read today, we can only surmise that Perckheyer's Greek original or some ancestor thereof, was written on a papyrus scroll, the medium of choice in Ptolemy's Alexandria. Not having invented paper yet, the Chinese were using silk cloth rolled and stored in boxes as well as bamboo strips, for writing media at that time. Parchment would not come into use for about 300 years. In India, palm leaves were inscribed in Sanskrit with a stylus, stained with turmeric and laced together into sheaves, a tradition that continued for 2,000 years, until the tradition fell into disuse in the nineteenth century. Today, as part of its role in the UDL, India is attempting to locate and preserve an estimated 100,000 inscribed palm leaf manuscripts, which, due to a variety of factors, are at risk of being physically lost to the vicissitudes of time, along with the knowledge they contain and the very languages in which they are written.

While digitizing centuries-old palm leaves and 16th-century translations of ancient texts presents no legal obstacles, digitizing more recent works is fraught with difficulties. Oddly, as a result of the triumph of the publishing industry...

2009 | 2010
DANCE

Subscribe today!
Call 412-456-1390
pgharts.org

Kidd Pivot
Saturday, October 3, 2009

Margaret Jenkins Dance Company & Guangdong Modern Dance Company
Saturday, October 24, 2009

Aspen Santa Fe Ballet
Friday, February 26, 2010

Complexions
Saturday, April 3, 2010

BJM Danse
Saturday, April 17, 2010

Vincent Dance Theatre
Saturday, May 1, 2010
Electricity/ knowledge

beginning in the early twentieth century, the vast majority of all the books ever written are currently either known to be in copyright or classified as "orphan books"—those of indeterminate copyright status. In either case, the situation makes digitizing them somewhat risky, if not outright illegal.

St. Clair explained, "Orphan books are those whose copyright ownership is not clear. Works published before 1923 are generally out of copyright. Books published between 1923 and 1963 had to be renewed in order to remain in copyright. Our estimate is that 80 percent of those books were not renewed, so they are not in copyright. But in order to determine that authoritatively, you have to go to the copyright office and do a manual search. So we are in the process of scanning and putting up some of the copyright renewal records."

In response to the difficulties posed by intellectual property issues, and in consideration of the fact that the publishing industry has not settled upon a way of trading in digital books, for the moment, the UDL has focused on scanning works that are clearly in the public domain, which are estimated to number about 3 million.

Finding the books to digitize is a matter of finding the libraries that own them. And that is largely a matter of associating with the people who run them. The CMU Library belongs to the Digital Library Federation, a group of 37 of the world's most prestigious libraries focused on realizing the promise of digital libraries. Although a benefactor occasionally purchases a library for scanning, most of the collections come from large academic libraries. Contrary to popular perception, the bulk of most libraries' collections is not in the public stacks but in storage, a fact that allows scanning to be done with minimal interruption to a library's routine services.

But not all librarians welcome the prospect of making their collections available online. "Libraries have attitudes," Shamos said. "Lots of large libraries tend to believe that their collections are related to their identity as a library, their budget and their ability to physically draw people into the doors."

Perhaps more lamentable than the narrow views of some librarians and curators is
the intentional concealment from public access of unique world treasures, a practice with which UDL's Shamos has had at least one disheartening experience.

"I got a meeting in New York with the curators of the manuscripts collection at the Morgan Library, which has one of the largest collections of illuminated manuscripts in the world," Shamos said.

"Because the manuscripts are so old, they are not in copyright. They sit in vaults and they're taken out about once every 50 years. I laid out the Universal Digital Library and what we were doing, and I invited them to participate.

Understanding the fragility and value of the works, I said, 'we will set up a scanner in the library so those manuscripts will never have to leave the room that they are in now.' At the end of my presentation they said very politely, 'Why would we ever want to do something like that?' I said, 'Well, this would be an easy way of disseminating the manuscripts.' To which they replied, 'What makes you think we have any interest in disseminating these things?' So I couldn't convince them to go along with us, and such is the attitude of many collections."

True to the UDL's enduring nature, Shamos counters the Morgan's rebuff with the knowledge that time is on his side. "I strongly believe in the embarrassment factor," he said. "We estimate that there are well over 3 million books that are out of copyright. All we have to do is find those and scan them. On the way to 100 million books, there's a lot of stuff that you can scan first. So when the No. 2 collection of illuminated manuscripts makes its collection available online, the Morgan Library will immediately change its position. I am never going to have to talk to them again. They'll come to us."

As challenging as locating public domain works may be, the next steps in the process—scanning, optical character recognition, indexing, translating and posting, all of which the project anticipates being done by machine—make them pale by comparison.

The linchpin of the digitization process is the overhead book scanner, which, unlike the commonplace flatbed
scanners on many desktops today is really a high-resolution digital camera suspended over a book stage, sort of like a drill press or giant microscope. The scanner images two pages of a book, compensating for page curvature and removing images of the operator’s fingers if he or she is holding the pages down. With some 50 or so scanning centers equipped with as many as 90 scanners, the process is sufficiently fast to permit the current scanning rate of 7,000 books per day.

Once the scan is complete, a TIFF file is sent to a computer for optical character recognition (OCR) and metadata harvesting. Because scanning can be done at a much faster rate than OCR, the processes are frequently performed asynchronously. At this stage of the project, the objective is to get the images scanned and get the books back to their place of origin as quickly as possible.

As ambitious as scanning and digitizing every published work of mankind may be, the age-old librarian’s question of how readers will be able to find what they are looking for remains a persistent problem. It is answered by the practice of metadata harvesting, in which the kind of information typically found in library catalogs is extracted from the title page of a book as soon as it is out of the OCR process. Unfortunately, the modern indexing systems that we take for granted had not yet been invented when many old books were written. So it is the responsibility of the person scanning the book to make certain that the metadata is as accurate as possible. Once the metadata is compiled, it is posted to the UDL index along with the book’s associated scans and text files. Today, the UDL posts files in HTML, PDF and TIFF formats.

Just as Reddy anticipated the advent of advanced technologies to facilitate his vision of electronic libraries a quarter-century ago, Carbonell, who also serves as director of CMU’s Language Technology Institute, anticipates the advancement of machine translation in the near future. “After indexing, we want to be able to do cross-language indexing,” Carbonell said. “After that, the next step would be a gist translation. We have the technology to do gist translations now, but we have not yet scaled it up to the universal library.” Balancing machine translation’s need for refinement with its
etty much the same—and still a brilliantly innovative example of residential re-planning and design.

In addition to their homes and commercial districts, Mt. Washingtonians are being protected and improving their green spaces, as well. In 2009, the beautiful Scenic Byway Park was dedicated for 280 acres ringing the mountaintop. Adjacent neighborhoods, and two years later a $154,000 federal grant was awarded to implement the master landscape plan. Prior to consolidation, these parks were a hodgepodge of mini-parks, run-forested hillsides, ballfields and a few scattered parcels. Now, thanks to the Mt. Washington Community Development Corporation, hiking and biking trails will connect the three parklets with the wooded areas to maximize the rural beauty and outdoor recreation offerings.

Pittsburghers are nothing if not sourceful, after all. Mayor Luke C. Walentas says the city is considering using windmills atop Mt. Washington to generate electricity with wind-driven turbines and thus cut energy costs. (He's a realist who acknowledges the challenging difficulties: “You know how easy it is even to build a building on Mt. Washington, much less a windmill there.”)

Symbolic of such new and old challenges is artist Jim West’s striking statue of George Washington and Seneca leader Guyasuta there in October 1770. They first met in 1753, when Guyasuta led a young officer, Washington, up the Allegheny River to serve notice on the Iroquois that they should vacate the area for the British. That order was declined, leading to the long, bitter French and Indian War, in which Washington and Guyasuta were first allies, then enemies. West's sculpture shows the old warriors in council 13 years later, discussing the peace of the seven nations' confederacy. Tension as well as respect are evident between the two leaders, who had diametrically different goals for the area. Washington had surveyed it for future white European settlement;

Guyasuta reminded him of the 1763 British treaty prohibiting just such expansion.

The perfect title of West’s work? It's called "Point of View." West says he hopes modern folks might visit it when they want to work out their differences—or just drink in the Point and the View ranked the second most beautiful in America by USA Weekend Magazine: “In a nation with a wealth of stunning cities full of compelling stories, ranking Pittsburgh as the No. 1 beauty spot is perhaps our most surprising choice. But the Steel City’s aesthetic appeal is undeniable, as is its very American capacity for renewal.”

Indeed, what could the No. 1 choice have been? Who cares. All we know and care about is that dazzling panorama of hills and valleys and skyscrapers nestled where theAllegheny and Mon form the mighty Ohio, while the lights twinkle up at us from the city and its bridges below.

The romance in your life seems to be waning? No need for Viagra when you have the view from Mt. Washington at night. 

Barry Paris is an award-winning biographer, film historian, Russian translator and magazine contributor.

Electrifying knowledge

CONTINUED FROM PAGE 92

increasingly rapid development, Carbonell anticipates application of the technology in the near future. “Right now we are at the cusp of robust machine translation. In two or three years it will start to become worth using.”

Happily, the Universal Digital Library’s recent success with The Million Book Project promises to provide both impetus and grist for the further development of the collection, as well as for the technologies that will enable the attainment of its objectives.

One of the project’s unexpected ancillary benefits may be the expansion of digital access around the world. “If one of the reasons for creating infrastructure is that every citizen of a country will have access to the same stuff that Harvard students do, it makes sense to build infrastructure and go broadband,” Shamos said. “We didn’t anticipate that the Universal Library itself might be a stimulating factor for IT infrastructure.”

Ultimately, Carbonell envisions the UDL as bringing four progressively intractable socioeconomic obstructions to the acquisition of knowledge: “The knowledge divide: Is all the world’s knowledge accessible to everyone everywhere in the world? The economic divide: Where is all the wealth that you can turn into knowledge? The infrastructure divide: Can anyone use this wealth? And the time divide: Can you turn that knowledge into a product?”

"That's the UDL in a sentence, it's the new infrastructure."
divide. Can everyone in the world afford access? The linguistic divide: Can everyone read it without having to learn another language? And the literacy divide: Can a person read at all?

Expressing a vision that pushes the UDLs mandate past simply making books available for reading. Carbonell anticipates the ongoing success of machine translation from OCR to text, and text across languages, as well as text-to-speech technology that will make information available to anybody anywhere, without regard for their ability to read.

St. Clair summed up the project's current state of affairs. "We did this on a shoestring, using a lot of good will and funding from a lot of different places. So now that we have our first success, we're looking for money to hire people to manage and run a larger operation."

For Reddy, the UDLs success is the realization of a dream. "It's what I call the democratization of knowledge, so that the people who have the books can keep them and the whole world can read them."

Assessing the project's progress to date, Shamos said, "In the beginning, many people laughed at the idea that we were going to do a million books when we didn't even have them all in one place. The first NSF grant didn't fit very neatly into their funding criteria, but now that we have this success we can start to solve problems that the NSF really likes to solve, such as: How do you navigate through huge corpora? How do you do digital preservation as media change? What's the right way to do persistent URLs? How do you detect duplicates? What do you do about copyright status? There's a zillion proposals that you could come up with for this. And it doesn't matter how long this takes," he said, "because getting there is half the fun.

Tom Imerito is president of Science Communications, a Pittsburgh technology communications consultancy.

Against the odds CONTINUED FROM PAGE 101

University of Pittsburgh (MHCID, later UPMC), headed by Detre with Romoff as executive vice president. It included WPIC, Falk Clinic, Presbyterian, Eye & Ear and the School of Medicine. Within 20 years, the WPIC model would drive UPMC to national and international distinction.

WINDING DOWN

KATHERINE DETRE constantly badgered her husband about his smoking. "Only you and the truck drivers still smoke," she told him. Detre, however, continued to puff away with an ebony cigarette holder at a three-pack-a-day rate. A heart attack in 1992 ended his smoking and altered his career. Romoff assumed responsibility for UPMC, and Detre remained as the university's senior vice chancellor for the health sciences. He finally stepped down in 1998 and Dr. Arthur Levine, a top director at the National Institute of Health, was recruited as senior vice chancellor for the health Sciences and dean of the School of Medicine.

The results speak for themselves. For the fiscal year ending June 30, 2008, system-wide revenues for UPMC were $7.1 billion. UPMC has 20 hospitals in Western Pennsylvania and expanding international operations in Italy, Ireland, the U.K. and Qatar. As of December 2008, the UPMC Health Plan insured 1.3 million lives, and UPMC has become an integrated global health enterprise and one of the leading nonprofit health systems in the nation.

Hand-in-hand with the rise of UPMC go Detre's transformation of the health sciences at the University of Pittsburgh. At the time of Detre's arrival, the university ranked 37th in NIH funding; when he stepped down in 1998 it stood 10th. In fiscal 2007, the health sciences received NIH grants of $457 million, making the university the 6th largest recipient of NIH research funding. Detre allows as how the move from 10th to 6th may have been more difficult than the long climb from 37th to 10th.

Dateline Pittsburgh—Jan. 14, 2009: "Pitt scientists make key diabetes finding. A team led by Dr. Andrew F. Stewart, chief of the university's division of endocrinology and metabolism, has identified key molecules..." When you see articles such as these in local and national papers—there have been hundreds in the past and there will be hundreds in the future—think of Tom Detre. They are ongoing testimony to his enduring contributions to our region's mental, physical and economic well-being.

The soon-to-be-released "Beyond the Bounds: A History of UPMC" by Mary Brigman aided the research of this report.

William S. Dietrich II is a trustee and chief investment officer of the Dietrich Charitable Trusts.