

Researchers Lead International Endeavor to Enhance Digital Libraries

BY ELIZABETH O. COOPER

Forget about the Dewey Decimal System and card catalogs. When it comes to federated digital libraries, Old Dominion University researchers are utilizing cutting-edge technology to connect Internet users with a vast array of information resources.

Digital libraries are becoming increasingly popular research areas for information retrieval of database methods and techniques. Universities, governmental agencies, scientific organizations and other groups are opting to collaboratively build and own a massive, centrally pooled library of data that users can access with just a few keyboard clicks. Four faculty members in Old Dominion's Department of Computer Sciences are lead-

ing an internationally recognized effort to construct and demonstrate novel digital library services. The group's work has been funded by various federal and state agencies, including the National Science Foundation, NASA, Los Alamos National Laboratory and the U.S. Navy.

"The most important problem facing digital libraries is that all digital libraries are islands," says Kurt Maly, chair of computer science, who is leading Old Dominion's digital library group along with Mohammad Zubair, professor of computer science. "Each digital library owns its software and user group and has its own purpose."

According to Maly, connecting these islands makes different information resources from anywhere on the World Wide Web more accessible to users. Enter the Open Archives Initiative. The OAI, which evolved from the need to increase access to scholarly publications by creating interoperable digital libraries, solves problems of digital library interoperability by defining simple protocols for information retrieval.

The Beginnings of OAI

Launched in 1998 at the Los Alamos National Laboratory, the core concept of OAI was initially known as the “Santa Fe Convention.” The Universal Preprint Service Prototype was developed under the lead of Old Dominion’s research team. The Prototype harvested nearly 200,000 records from several different archives and created an attractive environment for users. “It was incredibly difficult to do, but we built it and it worked,” recalls Maly.

The OAI divided the world into data providers and service providers. Data providers consisted of digital libraries that expose metadata (which is basically data about data), including a topic’s author, title and other crucial information that identifies the material to all who wish to access it. Service providers harvest metadata, putting the information together to provide a service, such as a search.

“To relate it to the physical world, I make copies of a catalog of cards and put them in one case so users can come to that place and look at any part of the book in which they are interested,” explains Zubair.

The group returned to Los Alamos in 1999, at which time the Santa Fe Convention was renamed the Open Archives Initiative, and an international movement to build bridges across islands of digital libraries took shape.

“OAI was born, and the movement really blossomed,” Maly recalls. “Libraries all over the world were participating. OAI does for digital libraries what the Internet did for islands of isolated networks.”

ARC – the Google of Digital Libraries

Xiaoming Liu, a Ph.D. student in computer science, Maly, Zubair and Michael Nelson, assistant professor in computer science, immediately set out to develop ARC, the first federated digital library which went online in 2001. One of the first federated search services based on OAI protocols, ARC harvests metadata from various OAI compliant archives, standardizes them and stores them in a search service. As a federation of over 160 libraries, ARC currently has more than 6 million metadata records.

“The contents are at the U.S. Library of Congress,” Maly says. “But the metadata is in Old Dominion’s service. We point to the contents.”

Zubair compares searching ARC to using the Google search engine to find Web pages. “What Google has done for the Web, we want to do for digital libraries.”

Maly adds that Google handles billions of Web sites. “We are probably three orders of magnitudes away from that.”

More than 160 digital libraries participate in ARC, including the Library of Congress, Virginia Tech, Humboldt-University of Berlin and ImageBase. All have joined OAI and are OAI compliant. “This is a living federation,” says Maly. “We go out daily and harvest the latest things published in the libraries. Every day there will be more articles. A year ago, we had 1.5 million records. This has become one of the major pieces of software that other universities use to establish smaller communities of libraries.”

He adds that other organizations that are not part of ARC have used software developed by Old Dominion computer scientists to create their own communities of federated digital libraries.

“Many digital libraries do not announce OAI compliance to everybody. They want to be a small community where everyone in the community uses it, but it’s not necessarily known to the world.”

Agencies can create a federation of digital libraries for their own interest group, such as ARCHON, a federated digital library for physics communities. The Old Dominion digital library group has been instrumental in developing ARCHON. “It has richer metadata,” says Maly. “We can do a lot more and provide higher services. It’s a nice thing we can do for specified libraries that we cannot do for general ones because there’s too much information in a general digital library.”

Along with ARC and ARCHON, Old Dominion’s digital library group claims DP9, a method in which Web crawlers and search engines, such as Google, harvest OAI repositories. Developed by Old Dominion computer science doctoral students, DP9 software issues commands to a digital library and presents the results in a format that the Web crawler can understand.

“It’s like a broker between the Web, the database and the crawler,” Maly explains. “When you search in Google, you are only searching in the shallow Web. The deep Web is hidden from all search engines. Anything in a database anywhere can be in the deep Web. Search engines can crawl around and find links to other Web pages with DP9 and can get a lot of information.”

Researchers seeking an abundance of data were the impetus for creating digital libraries. The first major funding for digital libraries came about in 1995 when the National Science Foundation funded the Digital Library Initiative. Along with the Carnegie Mellon Foundation, the NSF provides the most funding for digital libraries. The National Science Digital Library is a major NSF program, which includes 70 projects and has a broad group of members. Old Dominion’s Digital Library Group oversees three of those projects, including ARCHON.

Directing Traffic in Digital Libraries

Although they began as a tool for academicians, digital libraries have recently been developed for the public. “Nowadays, the digital library is really meant for everyone,” Maly says.

To direct traffic in digital libraries and ensure searches are orderly, Nelson is working with Buckets. These intelligent, mobile, data structures transfer the responsibility of preservation and content maintenance from the archive and the archivist to the object in question. Buckets protect, manage and mobilize a library’s content and basic services. “They are small, portable digital libraries that can be moved around and can communicate with each other,” Maly notes.

In addition, Old Dominion has received a \$300,000 two-year NSF grant to build personalized, self-sustainable, digital libraries. A traditional digital library is based on a centralized framework, which requires an organization in the community to take the lead in providing the hardware and software infrastructure and developing processes to maintain the content. The Old Dominion group is investigating alternate models for digital libraries based on peer-to-peer networks. These networks are decentralized, distributed and autonomous, which support evolution of communities from the bottom up – a similarity to evolution of communities in social networks.

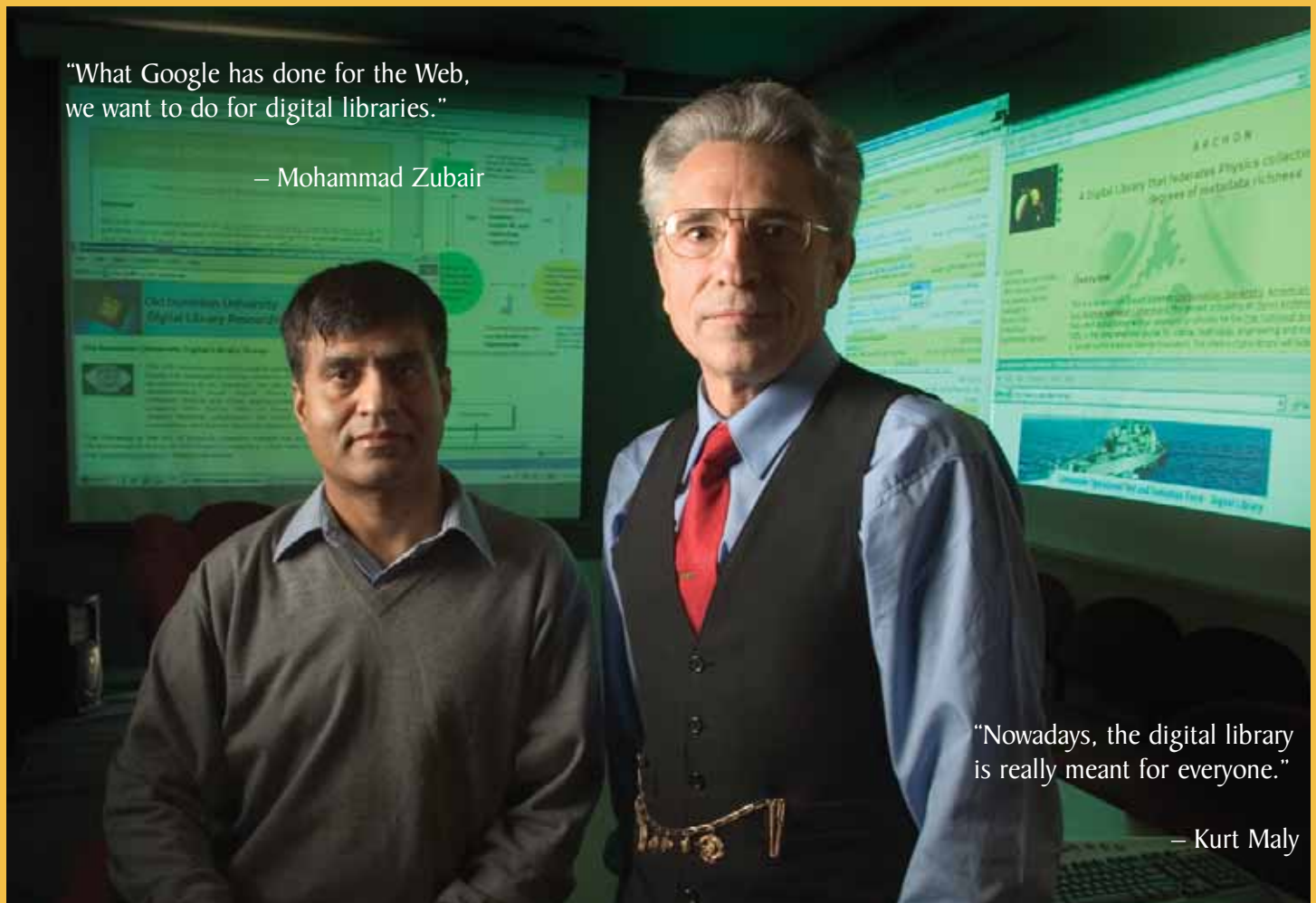
“You search for whatever people have in the digital library,” Maly explains. “People join the peer-to-peer network on their own. The software brings people with common interests together.”

Zubair adds that anyone can join the peer-to-peer network for free. “Everyone just runs it themselves. If you publish, you have to put metadata in it. Each one has to invest an effort to make the whole thing work. So much content in digital libraries is not being organized. That will be where the peer-to-peer network plays a role.”

Maly adds that much of this content exists on organizations’ Web sites, but they usually do not include metadata. “Metadata is what makes searches successful,” he adds. “Everyone does a little bit of work, and the whole community benefits. You get what you need and fast.”

However, despite the growing popularity of digital libraries, not everyone is jumping on the bandwagon. “Some commercial publishers refuse to be part of a federated digital library,” says Maly. “It’s a little bit of a contest to them rather than a collaboration.”

Still, he adds that digital libraries are quickly becoming a force to reckon with on the Internet, with all of them based on OAI compliance. “This will become the future. It’s already evolved tremendously in just a few years.”



“What Google has done for the Web,
we want to do for digital libraries.”

– Mohammad Zubair

“Nowadays, the digital library
is really meant for everyone.”

– Kurt Maly