Protecting Our Legacy Document Infrastructure: Threats and Opportunities

Testimony given by Daniel Lopresti
Class of 1961 Associate Professor
Department of Computer Science and Engineering
Lehigh University

The Pennsylvania House of Representatives
Veterans Affairs & Emergency Preparedness Committee
Public Hearing - Disaster Preparedness

Thursday, December 1, 2005, 9:30 a.m.
Lehigh University, Tower Room, Iacocca Hall, Bethlehem, PA

Good morning and thank you for giving me this opportunity to tell you about a problem I see that could have serious implications for Pennsylvania municipalities, businesses, and citizens. By way of background, I will note that in addition to serving on the faculty at Lehigh, I am known internationally for my research in the fields of pattern recognition, document analysis, digital libraries, and cybersecurity.

When we talk about critical infrastructure, it is natural to think of bridges and highways, hospitals, and the equipment used by public safety and health professionals. In recent years, it has become clear that our cyber infrastructure is also vital. To a large degree, this is because we are becoming a society more and more dependent on information and its flow. But there is a much older information infrastructure that is just as important, remains just as pervasive (if not more so), and is probably just as much at risk. That is, hardcopy or legacy documents. Each of us depends on this "other" information infrastructure every day. It is integral to our economic and physical well-being.

At one point, there was a belief that we would someday see the "paperless" office. This predication has proved to be far too optimistic – or misguided – or both. The fact is, paper still has tremendous ergonomic advantages over existing computer display

technologies. It also has some archival advantages over digital media as well. Hence, this should not be regarded as an "either / or" type of scenario. Electronic and paper documents should be allowed to continue to coexist. The challenge is getting them to work together.

One major difference between electronic and paper documents is the ease with which backups can be made and archived to protect the information contained within the document. You have certainly been told about the importance of making regular backups of your computer files, and hopefully you are taking such precautions to protect your data seriously. Of course, the safest backups are ones that are stored off-site, so that the kinds of catastrophes we see in the case of events like Hurricane Katrina do not wipe out the only copies of our files. However, my belief is that we employ far less effective strategies for protecting key legacy documents. Instead of being proactive, we are forced to react after-the-fact when disaster strikes.

I just want to take a moment to read from several news stories reported in the national media to hammer this point home.

This is from the Los Angeles Times, dated September 9th of this year:

Lives and history adrift on a soggy paper trail

Paper is everywhere in New Orleans — floating in the water, trapped in tree branches, ground into curbside mud. Millions of pages are soaked in courthouse basements, businesses and homes. Among the items are records of families, land ownership and commercial transactions, along with all the paper that charts the minutiae of everyday life.

In the basement of the Civil District Courthouse, water has lapped over 20% of the 60,000 leather-bound books of the New Orleans Notarial Archives. The books contain the records of all property transfers in the city that have occurred in the modern era.

"We don't have deeds in New Orleans," said Stephen P. Bruno, custodian of the archives. "Whatever our records say, that's who owns the property." ...

My understanding is that it will be a year or longer before it becomes possible to sell real estate again in some parts of New Orleans because of the loss of these records.

From National Public Radio on September 13th:

Many Medical Records Lost in Katrina Flooding

Victims of Katrina face more headaches: some evacuated but their medical records did not, and they may have been destroyed.

From the Chicago Tribune on the same date:

Katrina Disaster Disrupts Louisiana Legal System

Hurricane Katrina uprooted half of all practicing attorneys in Louisiana and upended the state and federal legal system ...

Evidence for an untold number of criminal investigations might be lost, along with records from hundreds of private law firms.

Records from the Court of Appeals may be under water ...

Going back a bit further to September 11th 2001, this is one of the many reports filed in the aftermath of those tragedies (from the New Civil Engineer website, dated September 27th, 2001):

Buried drawings hampered rescue efforts

Engineers were unable to call on vital engineering drawings of the World Trade Center for three days because the originals were kept in one of the collapsed twin towers ...

Without engineering drawings of the six story basement, engineers had to guess whether debris would support cranes and whether heavily damaged buildings would stay standing ...

"When you are guessing you are more cautious. Having those drawings straight away would have saved a lot of time, " said one of the engineers at the disaster site.

One obvious question, then, is how big is the problem of protecting our legacy document infrastructure? In all honesty, I cannot provide you with a concrete answer to that question this morning, but from what we are reading with regard to Katrina, it appears to be a bigger problem than anyone anticipated. It certainly seems like a problem worth studying.

What is the solution, then? At its plainest, the answer to that question is the same one you know for your computer files: replication, i.e., backups. These could be hardcopy backups, or, better yet, electronic backups with all their attendant advantages.

There are, of course, document conversion and scanning services, and many large banks and insurance companies already image their customer records. In Bethlehem, just a couple miles from where we sit, the OCLC Preservation Service Center performs archival imaging for a number of libraries around the country, including the Library of Congress. So far as I know, however, no cost-effective solution currently exists for municipalities, small businesses, and private citizens. And this is where my expertise in the field of document image analysis comes into play. It is one thing to scan a page of information to save it away someplace where you can hopefully find it in the future, and another thing entirely to scan a piece of paper and raise it fully and completely into the world of electronic documents. That requires solving some very hard problems in pattern recognition, problems that we humans have evolved to be quite good at, but where machines still fall woefully short. You have no difficulty reading text under a wide variety of conditions – different lighting, font-sizes and typefaces, rotated at a angle, on paper that is torn or coffee-stained or crumbled into a ball or photocopied too light or too dark. But current computer algorithms have a lot of trouble reading text like this accurately and often breakdown completely when confronted by noisy input.

You are no doubt aware that google is many people's current darling high-tech company. Google is making oodles of money indexing billions of web pages and delivering them to its users. Google's success is based on an intriguing premise: that search, which can be automated, trumps filing, which is inherently manual. If you can find a way to search billions of documents, you can completely dispense with filing, a much more time-consuming and expensive task. This is one of the big hurdles, I believe, in fully automating the capture and archiving of legacy documents, so we can learn a valuable lesson by watching what google is doing.

It is also an interesting coincidence that google, having indexed the “easy” part of the

Web, is now applying its tremendous resources to a tougher challenge, scanning more or less every book in the libraries at Harvard, Stanford, Oxford, the University of Michigan, and the New York Public Library. Take note of the google approach because I think it is instructive: scan everything. Do not bother wasting time deciding what is important and what is not – let the software sort that out later.

I should note that, when it comes to document image analysis, google is solving a relatively easy problem. The books in question are most often carefully produced and well maintained – the text on those page images is very clean, so character recognition accuracies are high – 98% or better. Real-world documents – the kinds you probably have sitting on your desk and that I was talking about earlier – are usually much more difficult to process automatically. For noisy inputs, we are often lucky to obtain 80% character recognition accuracy, and that is enough to break search techniques like google's. Unconstrained handwriting recognition, which is even harder than recognizing typeset text, is still considered an open problem.

I would be remiss if I did not end by mentioning that Lehigh is one the few American universities with research expertise in document image analysis as it applies to the problem of legacy documents. Both I and my distinguished colleague Professor Henry Baird are very active in the field. For example, we are currently working with several researchers elsewhere on a project for DARPA, the Defense Advanced Research Projects Agency, to survey the state of the art in document analysis systems, with an eye toward supporting the military and their needs. We are also investigating new techniques to improve character recognition accuracy in the case of degraded inputs.

Thank you for your time and I would be happy to try to answer your questions.