Alan Pierce
pierceaj@optonline.net

From Printed Books to Digital Searchable Files

Past technological achievements are the foundation of our modern day society. However, when many of these scientific discoveries and technological inventions were first introduced, they were seen as, to say the least, disruptive by the people who they displaced. There is no doubt that Gutenberg's 15th century printing inventions changed the availability of the printed word. However, the spread of his inventions was very disruptive to the scribes who were at the time duplicators of manuscripts and books.

The technologies that have brought us the printed word for the past six centuries are now being usurped by the digital-age concept of a paperless society. On December 14, 2004, Google announced that "the libraries of Harvard, Stanford, the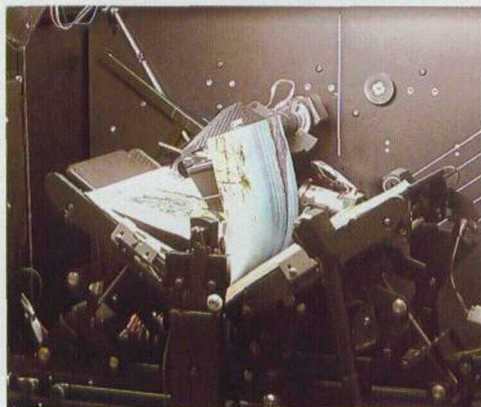 University of Michigan, the University of Oxford and the New York Public Library [will] join with Google to digitally scan library books and make them searchable online" (Google Press Center).

Larry Page, cofounder of Google, indicates that this undertaking is part of the Google Print Program that intends, over time, to make it possible for people to search online for specific information from the full content of most printed books.

The Association of American Publishers (AAP), at the request of five major book publishers, has started legal action to prevent Google from committing what they see as copyright infringement. However, Google intends to only provide a few sentences at a time from a book in response to a specific question. Therefore, they contend, even if they scan books that are copyrighted, they are protected by the fair-use doctrine of the Copyright Act. The situation is similar to writing a term paper and quoting and citing a specific book. You are not infringing on the book's copyright because of this same fair-use provision.

The legal issues will be solved over time. The question for this column is what technologies are used to quickly scan entire books and then convert their contents into a searchable format? The Kirtas APT Bookscan 1200 is a machine designed to automatically scan and digitize books at a speed of up to 1,200 pages per hour.

A book to be scanned is first placed in the bed of the machine. When the machine is activated, a set of vacuum fingers gently lifts a page, then a machine arm turns the page.

The arm carries the page over to the opposite side of the book to expose the next two pages for scanning. The scanner is located at the top of the machine and uses mirrors to bring the left and right page into position for scanning. Software similar to that found in digital cameras straightens any imperfections in the scanned image. Each scanned page is automatically written to a file. You can see a video of this machine in action at www.kirtastech.com.

At this point, the pages of the book are merely digital representations; individual words or sentences cannot be searched for meaning. Optical character recognition (OCR) software is then used to automatically process these scanned pages into search engine-recognizable text that you can cut and paste from your computer screen into a document. Imagine a factory floor full of similar machines digesting the written word at thousands of books per hour.

Photos courtesy of Kirtas Technologies Inc.

Above, the Kirtas APT Bookscan 1200

(At left) The machine's arm gently turns the pages of a book

Recalling the Facts

1. Book publishers feel threatened by the technology described in this column. How is this threat similar to what the scribes faced in the 15th century? How is it different?

2. What legal argument are publishers using to stop Google?

3. A scanned document can't be electronically searched. Why?

4. What kind of software can you use to convert your scanned document into individual words? ©

Alan Pierce, Ed.D., CSIT, is a technology education consultant. Visit www.technologytoday.us for past columns and teacher resources.