Diane L. Richard

reCAPTCHA: Deciphering Old Texts

Diane L. Richard investigates how security software is helping digitize documents and books!

I SPEND SO MUCH time looking into new resources, it's rare that I come across something that really impresses me. But reCAPTCHA, *http://recaptcha.net*, is one of the few applications that stood out.

If you are familiar with the ongoing project to digitize the *New York Times* and the Internet Archive's book digitization project, *www.archive.org*, you have already benefited from reCAPTCHA!

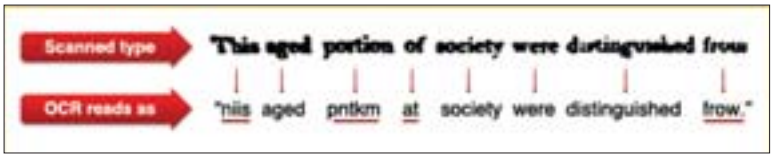
But how?

These and other projects are currently digitizing physical books or newspapers of historical, and genealogical, value. The pages are digitally scanned, and then transformed into text using Optical Character Recognition (OCR). However, OCR is not perfect.

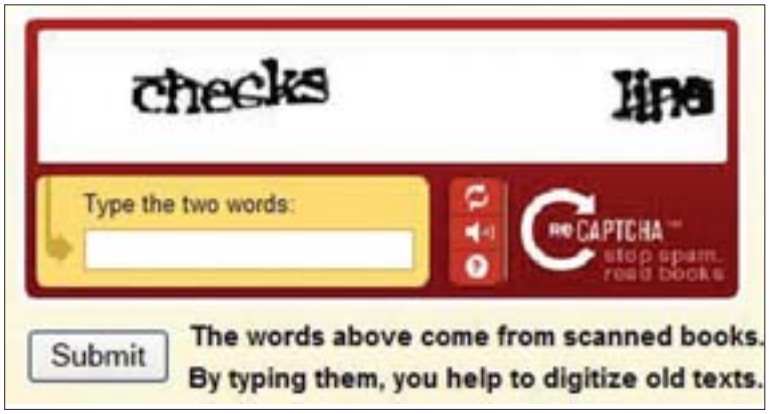
Anyone who has tried to use OCR, or seen the results of it, is familiar with how powerful a tool it is, and yet how imperfect the outcome can be. If a page is skewed, has blots on it, the type is hard to distinguish or contains any of a number of other flaws, this increases the likelihood that OCR will not be able to perfectly convert the text in an image to a pure text version.

ReCAPTCHA improves the process of digitizing books by isolating words that can't be read by computers for humans to decipher. This is possible because most OCR programs alert you when a word can't be read correctly. Each word that cannot be read correctly by OCR is used as a CAPTCHA.

Once they are further distorted using lines and warps, these CAPTCHAs are placed in front of consumers. They are used as security words on websites to make sure that you are a human (versus another computer) performing the transaction! Whereas a human can interpret and retype the displayed security word (CAPTCHAs), a computer can't.



Above: This example shows what can get lost in translation when using OCR. Words that are legible to most people are sometimes indecipherable to the program. Below: This security word example from reCAPTCHA shows you what you might see on a website that uses this technology.



You have probably come across this technology if you have made an online purchase, signed up for a mailing list or accessed a secure website. More than 40,000 websites — including popular ones such as Ticketmaster, Facebook and Craigslist — use reCAPTCHA; so, you may have already helped digitize a newspaper or book and you didn't even know it! More than 60 million CAPTCHAs are solved every day by people around the world.

Developed by a computer scientist at Carnegie Mellon University in Pittsburgh, it is offered for free through the university, and has already digitized more than 13 billion words. It has allowed the *New York Times* to completely digitize about two years' worth of newspapers every month!

And, don't worry if you can't "read" the word — the image will actually be shown to several people who have to agree on what the word is before it will be considered accurately transcribed. And, since the OCR software can't always differentiate between an illegible word and an ink blot, sometimes what you are offered truly is a "blob", and not a distorted word.

Now, the next time you type a CAPTCHA on a website, you just might have helped digitize a document that will benefit your genealogy research!

DFH