Programas de OCR auxiliam digitalização de documentos

A informatização da Justiça chegou a um impasse. Escritórios de advocacia, tribunais e departamentos jurídicos têm, entre suas principais dificuldades, a digitalização de toneladas de papéis que estão em seus arquivos. Milhares documentos mofam nas prateleiras ou são devorados por traças.

O problema é que não basta escanear as cópias para preservar o saber jurídico de décadas contido nessas páginas. Os arquivos que chegam ao computador pelo scanner são apenas imagens ou estão configuradas no modo PDF. O texto não pode ser editado. O processo de preservação torna-se parcial.

A questão envolve ainda documentos presentes no dia-a-dia do operador de Direito. Já é hábito dos advogados irem ao fórum com uma máquina digital em mãos para fotografar os autos. Quando chega ao escritório, ele apenas pode consultar o documento, sem poder manipular o texto escrito.

Para esse tipo de trabalho, no entanto, existem os programas de OCR (sigla em inglês de Reconhecimento Óptico de Caractere). O mercado dispõe de dezenas de opções. Alguns equipamentos como scanners e multifuncionais trazem esse tipo de aplicativo de graça. Porém, o resultado nem sempre é satisfatório. Em alguns casos, é mais fácil digitar o texto do que usar o OCR.

As soluções mais profissionais têm resultados melhores. Os programas mais conhecidos são o FineReader, da empresa russa Abby, e o Omnipage, da norte-americana Nuance. O primeiro está na sua versão 9.0 e foi lançado recentemente no Brasil. Ele pode ser comprado pelo site da empresa ao preço de US$ 250. O outro também está disponível no site. A versão 16, lançada no final do ano passado, sai por US$ 400.

A **Consultor Jurídico** testou o FineReader com três textos de decisões de tribunais. Na primeira tentativa, feita com um documento datilografado do Supremo Tribunal Federal de 1946, o resultado foi ruim (veja o original **aqui** e o resultado **aqui**). Pelo que se vê, numa situação como essa, leva-se muito menos tempo digitando o texto do que consertando os erros.

Os outros testes, no entanto, mostraram que a qualidade da imagem influencia diretamente no resultado. Ao converter um arquivo tipo PDF do Tribunal Regional do Trabalho do Paraná, o produto final ficou excelente. Das 8.193 palavras, o corretor ortográfico do Word mostrou problemas em apenas 14 (veja o original **aqui** e o resultado **aqui**). No terceiro teste, com uma qualidade de imagem ainda maior, só houve problema em apenas uma das 1.584 palavras (veja o original **aqui** e o resultado **aqui**).

Quando for comprar um programa, o usuário deve ter em mente que um dos requisitos mais importante é ter a versão em português — os dois programas já foram traduzidos. A falta do recurso fará que o programa não reconheça os caracteres acentuados.

Também é essencial que o software mantenha a formatação do texto e o deixe organizado como está. No teste feito com o FineReader, os símbolos do Poder Judiciário ficaram como imagens, as quebras de páginas foram devidamente sinalizadas e os recuos do juiz permaneceram como foram feitos. A análise de linguagem é outro recurso fundamental para o bom desempenho do programa. Ele compara o texto

com as palavras da língua de forma a evitar erros bobos como a troca da letra L pelo número 1.

A Abby, empresa que foi criada ainda na época da União Soviética, elenca, como recurso importante da nova versão, a ADRTTM (sigla em inglês para Tecnologia Adaptável de Reconhecimento de Documento). A função faz o programa ver a estrutura do documento como um objeto completo. Também é citada pela a empresa a função que reconhece arquivos de câmeras digitais. Já o OmniPage 16 promete um resultado superior a 99% e uma formatação final “perfeita”. Os dois programas funcionam em Macintosh, Linux e Windows Vista e XP.

Date Created
23/03/2008