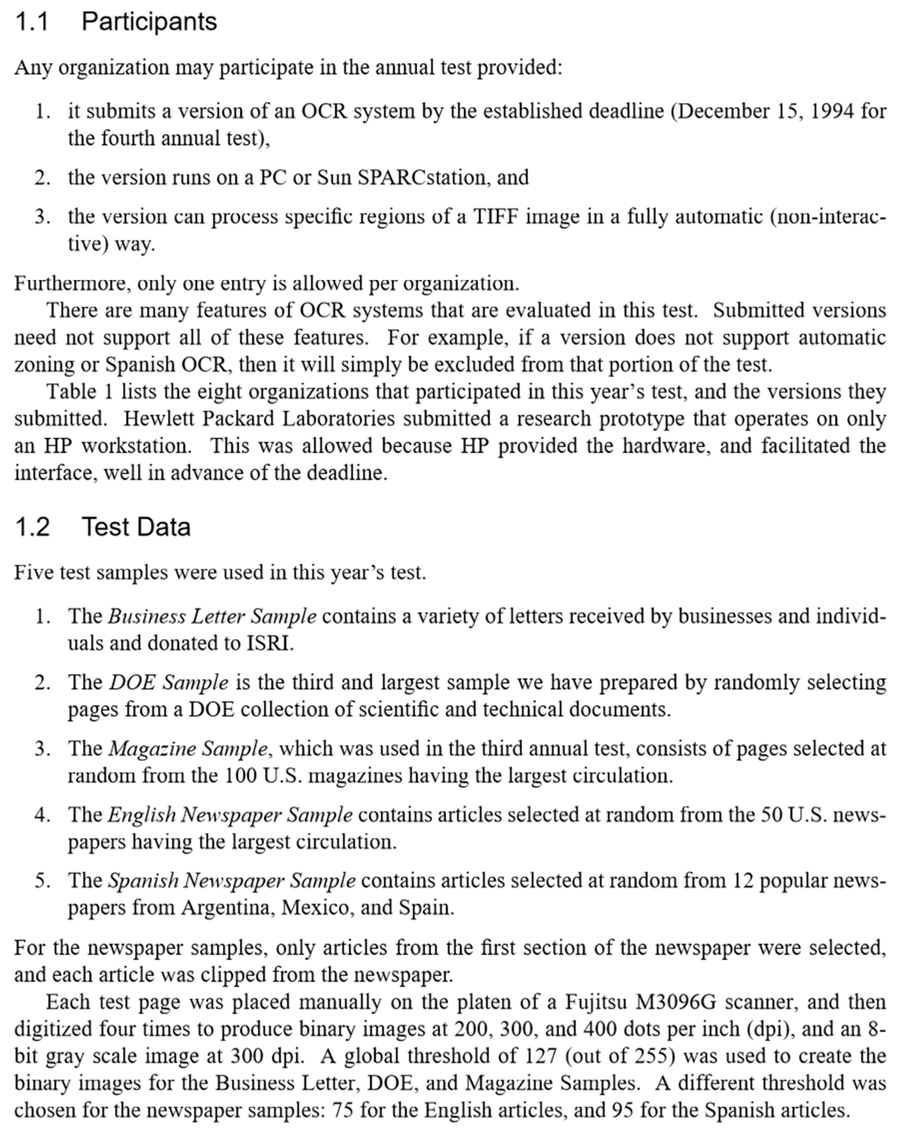
This is a Word document converted to PDF containing BEGIN_TEXT, END_TEXT and RENDER_IMG. The following is an image containing text that can be extracted with Tesseract.



1.1 Participants

Any organization may participate in the annual test provided:

1. it submits a version of an OCR system by the established deadline (December 15, 1994 for the fourth annual test),
2. the version runs on a PC or Sun SPARCstation, and
3. the version can process specific regions of a TIFF image in a fully automatic (non-interactive) way.

Furthermore, only one entry is allowed per organization.

There are many features of OCR systems that are evaluated in this test. Submitted versions need not support all of these features. For example, if a version does not support automatic zoning or Spanish OCR, then it will simply be excluded from that portion of the test.

Table 1 lists the eight organizations that participated in this year's test, and the versions they submitted. Hewlett Packard Laboratories submitted a research prototype that operates on only an HP workstation. This was allowed because HP provided the hardware, and facilitated the interface, well in advance of the deadline.

1.2 Test Data

Five test samples were used in this year's test.

1. The *Business Letter Sample* contains a variety of letters received by businesses and individuals and donated to ISRI.
2. The *DOE Sample* is the third and largest sample we have prepared by randomly selecting pages from a DOE collection of scientific and technical documents.
3. The *Magazine Sample*, which was used in the third annual test, consists of pages selected at random from the 100 U.S. magazines having the largest circulation.
4. The *English Newspaper Sample* contains articles selected at random from the 50 U.S. newspapers having the largest circulation.
5. The *Spanish Newspaper Sample* contains articles selected at random from 12 popular newspapers from Argentina, Mexico, and Spain.

For the newspaper samples, only articles from the first section of the newspaper were selected, and each article was clipped from the newspaper.

Each test page was placed manually on the platen of a Fujitsu M3096G scanner, and then digitized four times to produce binary images at 200, 300, and 400 dots per inch (dpi), and an 8-bit gray scale image at 300 dpi. A global threshold of 127 (out of 255) was used to create the binary images for the Business Letter, DOE, and Magazine Samples. A different threshold was chosen for the newspaper samples: 75 for the English articles, and 95 for the Spanish articles.