

# A Synthetic Scanned Document

This document stands in for a scanned magazine article from 1998 and has the same structure: five pages, each a black and white image 2,560 pixels across, drawn on a Letter page by two content streams. It was made for the tests and examples of PDFjet by util/make-scanned-pdf.py.

The pages that follow describe how the file is built, one feature on each page, so that a reader of PDF files meets what a file of this age can hold.

This is a test!

# 1. Linearization

The file is linearized, or optimized for fast web view. It starts with a linearization dictionary that gives the length of the file, the object number of the first page and the offset where that page ends, followed by a cross-reference section for the objects of the first page and a hint stream.

The cross-reference section for the other objects is at the end of the file, and the trailer at the top points to it with /Prev. The startxref at the end of the file points back to the section at the top.

This is a test!

## 2. CCITT Group 4 Images

Every page is a scanned image with one bit per pixel, compressed with the CCITT Group 4 fax encoding, /K -1 in its decode parameters. A page of text takes about ten kilobytes this way.

The image is drawn 547.2 points wide, which is about 337 pixels per inch, and the scans are a few rows taller or shorter than one another, as the pages of a real scan are.

This is a test!

### 3. LZW Content Streams

The contents of every page are an array of two streams compressed with LZW, the filter that PDF used before Flate. The first stream saves the graphics state, scales the image to the page and draws it; the second restores the graphics state.

Each page names its image in its own resources and shares one procedure set array, [/PDF /ImageB], with the other pages.

This is a test!

## 4. Old Habits

The lines of the file end in a carriage return alone, but for the entries of the cross-reference tables, which are 20 bytes long and end in a carriage return and a line feed.

The length of every stream is an indirect object written after the stream, the header says PDF 1.2, and the catalog opens the first page at the width of the window.

This is a test!