

Internal memo

To: Quality Assurance team
From: Sample Files, Technology Channel
Date: 12 February 2026
Subject: Scanned document for OCR testing

Text in a PDF is not stored as flowing paragraphs. Each piece of text is placed at exact coordinates on the page, often a few words at a time. Extracting readable text therefore means reassembling lines, columns and reading order, which is harder than it sounds. Multi-column layouts, tables and footnotes are classic sources of extraction bugs, and they are worth including in any test set.

Start with limits. Try a file just under the maximum size, exactly at it and just over it. Try an empty file, a file with no extension, a file whose extension does not match its contents and a file whose name contains spaces, accents or emoji. Check what happens when a user uploads the same file twice, or cancels halfway through.

Stable URLs and checksums make this reliable. If a file changes unexpectedly, the checksum comparison fails loudly instead of letting a test pass for the wrong reason. Keep the files small where possible so the suite stays fast, and reserve the largest samples for scheduled performance runs.

Signed,



Technology Channel



Test results

Audio bugs often hide in the details: an unexpected sample rate, a mono file where stereo was assumed, 24-bit samples in a pipeline built for 16-bit, or a long ID3 tag with cover art that pushes the first audio frame further into the file than a naive parser expects.

Case	Sample file	Size	Result
TC-001	sample-4032x3024.heic	1.3 MB	Pass
TC-002	sample-100-rows.csv	1.7 MB	Fail
TC-003	messy-folders.zip	4.5 MB	Pass
TC-004	sample-10-pages.docx	8.5 MB	Pass
TC-005	sample.xml	6.4 MB	Pass
TC-006	sample.txt	6.9 MB	Pass
TC-007	sample.json	1.4 MB	Pass
TC-008	sample-30s.mp3	5.0 MB	Fail
TC-009	sample.xml	4.5 MB	Fail
TC-010	sample.xml	7.7 MB	Fail
TC-011	sample.xml	8.4 MB	Pass
TC-012	sample-transparent-512x512.png	6.4 MB	Fail