Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

This document was generated on 27 November 2026 as a fixture for a PDF compression test. It is not an official record.

A PDF can hold selectable text, photographs, vector lines, or a picture of a page. Those ingredients do not shrink in the same way when you ask a compressor to cut the file down to a size cap.

Text stored as type usually stays sharp after a rewrite, because the letters are shapes, not a photograph of letters. Photographs stored as JPEG already threw some detail away when they were first saved.

A scanned PDF is different again: the whole page is a picture, so any further JPEG compression or downsampling hits both the words and the photos on that page.

The point of this fixture is to give those three cases a common starting point, with known photographs and a text layer we can extract later.

Page size is A4 (210 by 297 millimetres). Photographs are credited on the page that shows them, and the licence files sit next to this script.

Compression tools sometimes promise a smaller file without saying whether the text is still selectable, or whether a photo has been reduced to 72 pixels per inch.

This sample text is long enough to fill several pages so that a compressor has a real text stream to copy or, if it rasterises the page, to destroy.

When you read the study, the interesting comparison is not only the file size. It is also the SSIM of a 150 DPI render of each page, and whether pdftotext still returns the words.

Caps in the study are counted in decimal bytes: 100 KB is 100,000 bytes and 1 MB is 1,000,000 bytes, the same definition used on Cuelark's size-capped tools.

Ghostscript presets such as /screen and /ebook are included as a baseline that many people already use from the command line.

A size cap is a different request: it asks how far quality has to fall before the file lands under a named limit, or whether it can land there at all.

If a source file is already under a cap, the honest result is to leave it alone. The measurement script records that case instead of recompressing it.

The two photographs are unchanged copies of the Wikimedia originals. They are not cropped or re-exported before being placed on the page.

A caption under each photograph names the photographer and the CC BY-SA 4.0 licence, which is required if the page is reused.

Further pages repeat and vary this material so the mixed and text documents have a stable word count we can compare after compression.

This document was generated on 27 November 2026 as a fixture for a PDF compression test. It is not an official record.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

A PDF can hold selectable text, photographs, vector lines, or a picture of a page. Those ingredients do not shrink in the same way when you ask a compressor to cut the file down to a size cap.

Text stored as type usually stays sharp after a rewrite, because the letters are shapes, not a photograph of letters. Photographs stored as JPEG already threw some detail away when they were first saved.

A scanned PDF is different again: the whole page is a picture, so any further JPEG compression or downsampling hits both the words and the photos on that page.

The point of this fixture is to give those three cases a common starting point, with known photographs and a text layer we can extract later.

Page size is A4 (210 by 297 millimetres). Photographs are credited on the page that shows them, and the licence files sit next to this script.

Compression tools sometimes promise a smaller file without saying whether the text is still selectable, or whether a photo has been reduced to 72 pixels per inch.

This sample text is long enough to fill several pages so that a compressor has a real text stream to copy or, if it rasterises the page, to destroy.

When you read the study, the interesting comparison is not only the file size. It is also the SSIM of a 150 DPI render of each page, and whether pdftotext still returns the words.

Caps in the study are counted in decimal bytes: 100 KB is 100,000 bytes and 1 MB is 1,000,000 bytes, the same definition used on Cuelark's size-capped tools.

Ghostscript presets such as /screen and /ebook are included as a baseline that many people already use from the command line.

A size cap is a different request: it asks how far quality has to fall before the file lands under a named limit, or whether it can land there at all.

If a source file is already under a cap, the honest result is to leave it alone. The measurement script records that case instead of recompressing it.

The two photographs are unchanged copies of the Wikimedia originals. They are not cropped or re-exported before being placed on the page.

A caption under each photograph names the photographer and the CC BY-SA 4.0 licence, which is required if the page is reused.

Further pages repeat and vary this material so the mixed and text documents have a stable word count we can compare after compression.

This document was generated on 27 November 2026 as a fixture for a PDF compression test. It is not an official record.

A PDF can hold selectable text, photographs, vector lines, or a picture of a page. Those ingredients do not shrink in the same way when you ask a compressor to cut the file down to a size cap.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text stored as type usually stays sharp after a rewrite, because the letters are shapes, not a photograph of letters. Photographs stored as JPEG already threw some detail away when they were first saved.

A scanned PDF is different again: the whole page is a picture, so any further JPEG compression or downsampling hits both the words and the photos on that page.

The point of this fixture is to give those three cases a common starting point, with known photographs and a text layer we can extract later.

Page size is A4 (210 by 297 millimetres). Photographs are credited on the page that shows them, and the licence files sit next to this script.

Compression tools sometimes promise a smaller file without saying whether the text is still selectable, or whether a photo has been reduced to 72 pixels per inch.

This sample text is long enough to fill several pages so that a compressor has a real text stream to copy or, if it rasterises the page, to destroy.

When you read the study, the interesting comparison is not only the file size. It is also the SSIM of a 150 DPI render of each page, and whether pdftotext still returns the words.

Caps in the study are counted in decimal bytes: 100 KB is 100,000 bytes and 1 MB is 1,000,000 bytes, the same definition used on Cuelark's size-capped tools.

Ghostscript presets such as /screen and /ebook are included as a baseline that many people already use from the command line.

A size cap is a different request: it asks how far quality has to fall before the file lands under a named limit, or whether it can land there at all.

If a source file is already under a cap, the honest result is to leave it alone. The measurement script records that case instead of recompressing it.

The two photographs are unchanged copies of the Wikimedia originals. They are not cropped or re-exported before being placed on the page.

A caption under each photograph names the photographer and the CC BY-SA 4.0 licence, which is required if the page is reused.

Further pages repeat and vary this material so the mixed and text documents have a stable word count we can compare after compression.

This document was generated on 27 November 2026 as a fixture for a PDF compression test. It is not an official record.

A PDF can hold selectable text, photographs, vector lines, or a picture of a page. Those ingredients do not shrink in the same way when you ask a compressor to cut the file down to a size cap.

Text stored as type usually stays sharp after a rewrite, because the letters are shapes, not a photograph of letters. Photographs stored as JPEG already threw some detail away when they were first saved.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

A scanned PDF is different again: the whole page is a picture, so any further JPEG compression or downsampling hits both the words and the photos on that page.

The point of this fixture is to give those three cases a common starting point, with known photographs and a text layer we can extract later.

Page size is A4 (210 by 297 millimetres). Photographs are credited on the page that shows them, and the licence files sit next to this script.

Compression tools sometimes promise a smaller file without saying whether the text is still selectable, or whether a photo has been reduced to 72 pixels per inch.

This sample text is long enough to fill several pages so that a compressor has a real text stream to copy or, if it rasterises the page, to destroy.

When you read the study, the interesting comparison is not only the file size. It is also the SSIM of a 150 DPI render of each page, and whether pdftotext still returns the words.

Caps in the study are counted in decimal bytes: 100 KB is 100,000 bytes and 1 MB is 1,000,000 bytes, the same definition used on Cuelark's size-capped tools.

Ghostscript presets such as /screen and /ebook are included as a baseline that many people already use from the command line.

A size cap is a different request: it asks how far quality has to fall before the file lands under a named limit, or whether it can land there at all.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.

Text fixture for a PDF compression test

Generated 27 November 2026. Selectable text, no photographs.