



Features

Request trial

[← Back to Blog](#)

## What's so hard about PDF text extraction?

by Bogdan on July 14, 2020

be too difficult. After all, the text is right there in front of our eyes and humans consume PDF content all the time with great success. Why would it be difficult to automatically extract the text data?

Turns out, much how [working with human names is difficult](#) due to numerous edge cases and incorrect assumptions, working with PDFs is difficult due to the extreme flexibility given by the PDF format.

The main problem is that PDF was never really designed as a data input format, but rather, it was designed as an output format giving fine grained control over the resulting document.

At its core, the PDF format consists of a stream of instructions describing how to draw on a page. In particular, text data isn't stored as paragraphs - or even words - but as characters which are painted at certain locations on the page. As a result, most of the content semantics are lost when a text or word document is converted to PDF - all the implied text structure is converted into an almost amorphous soup of characters floating on pages.

As part of building [FilingDB](#), we've extracted text data from tens of thousands of PDF documents. In the process, we have seen how every single assumption we had about how PDF files are structured was proven incorrect. Our mission was particularly difficult as we had to process PDF documents coming from a variety of sources, with wildly different styling, typesetting and presentation choices.

The list below documents some of the ways PDF files have made it difficult (or even impossible) to extract text contents.

## PDF read protection

You may have come across PDF files which refuse to let you copy their text content. For example, here is what SumatraPDF shows when attempting to copy text from a copy-protected document.



do not copy this text

Interestingly, the text is already visible, yet the PDF viewer is refusing to populate the clipboard with the highlighted text.

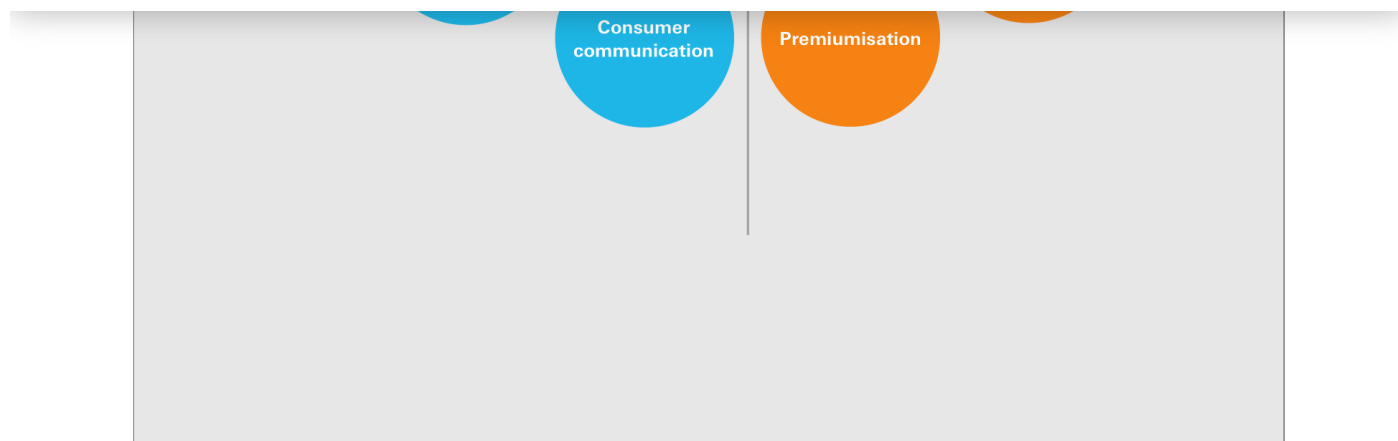
The way this is implemented is by having several “access permissions” flags, one of which controls whether copying content is allowed. It’s important to keep in mind that this restriction is not enforced by the PDF file - the actual PDF contents are unaffected and it is up to the pdf renderer to honour this flag.

Needless to say, this offers no real protection against extracting the text out of the PDF, as any reasonably sophisticated PDF handling library will allow the user to either toggle the flags or ignore them.

## Off-page characters

It is not uncommon for PDF files to contain more textual data than is actually displayed on the page. Take this page from the 2010 Nestle annual report.





There is more text associated with this page than meets the eye. In particular, the following can be found in the content data associated with this page:

“KitKat celebrated its 75th anniversary in 2010 but remains young and in touch with trends, having over 2.5 million Facebook fans. It is sold in over 70 countries and enjoys good growth in the developed world and emerging markets, such as the Middle East, India and Russia. Japan is its second biggest market.”

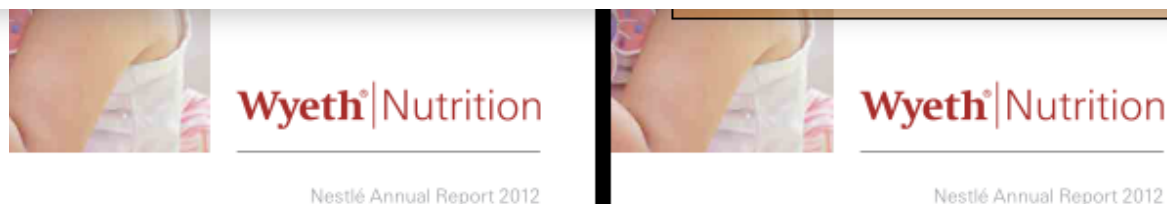
This text is actually positioned outside the page’s bounding box, so it is not displayed by most PDF viewers, but the data is there and will appear when programmatically extracting the text.

This occasionally happens due to last minute decisions to remove or replace text during the type setting process.

## Small / invisible characters on page

PDFs occasionally introduce very small or hidden text on the page. For example, here is a page from the Nestle 2012 annual report.





The page contains small white text on white background with the following contents:

"Wyeth Nutrition logo Identity Guidance to markets

Vevey Octobre 2012 RCC/CI&D"

This is sometimes done for the benefit of accessibility, similar to how the alt attribute is used in HTML.

## Too many spaces

Sometimes PDFs include extra spaces between letters in a word. This is most likely done for kerning purposes. ("Kerning" is the process of adjusting distances between characters during the type setting process)

**Example:** the 2013 Hikma Pharma annual report contains the following text:

CHAIRMAN'S STATEMENT

AN EXCELLENT

Copying the text gives:

"ch a i r m a n ' s s t a t e m e n t"

Reconstructing the original text is a difficult problem to solve generally. Our most successful approach has been applying OCR techniques.

## Not enough spaces

Character.

**Example 1:** The following extract from the 2017 SEB annual report.

## Global trends drive change

Ten years after the financial crisis started to spread, global trends have reshaped the banking industry. A new regulatory frame-

The extracted text shows:

"Tenyyearsafterthefinancialcrisisstarted"

**Example 2:** The 2013 Eurobank annual report shows the following

On April 7, 2013, the competent authorities deci  
Eurobank, which had been completely deprived of

Extracting the text gives:

"On\_April\_7,\_2013,\_the\_competent\_authorities"

Again, our most successful solution was to run OCR on these pages.

## Embedded fonts

PDF font handling is complex to say the least. To understand how PDF files store text data we must first know about glyphs, glyph names, fonts.

- A glyph is a set of instructions describing how to draw a symbol or character.
- A glyph name is the name associated with that glyph. For example "trademark" for the "™" glyph and "a" for the "a" glyph.
- Fonts are lists of glyphs with associated glyph names. For example, most

In a PDF, the characters are stored as numbers, called “codepoints”. To decide what to draw on the screen, a renderer has to go:

codepoint -> glyph name -> glyph

For example, a PDF document can contain codepoint 116, which it maps into the glyph name “t” which, in turn, maps into the glyph describing how to draw “t” on the screen.

Codepoint

116	104	101
-----	-----	-----

Font glyph name

“t”	“h”	“e”
-----	-----	-----

Font glyph

t	h	e
---	---	---

Now, most PDF files use a standard codepoint encoding. A codepoint encoding is a set of rules that assign meaning to the codepoints themselves. For example:

- ASCII and Unicode both use codepoint 116 to represent the letter “t”.
- Unicode maps codepoint 9786 to “white smiley face”, rendered as ☺, whereas ASCII is not defined at that codepoint.

However, PDF documents occasionally use their own custom encoding together with custom fonts. It might seem strange, but a document can use codepoint 1 to represent the letter “t”. It will map codepoint 1 into the glyph name “c1”, which will map into a glyph describing how to draw the letter “t”.

Codepoint

1	2	3
---	---	---

## Font glyph

t	h	e
---	---	---

While for a human the end result looks the same, a machine will get confused by the codepoints it is seeing. If the codepoints do not follow a standard encoding, then it is virtually impossible to programmatically know what codepoints 1, 2 and 3 represent.

Why would a PDF document contain nonstandard fonts and encodings?

- One reason is to make text extraction more difficult.
- Another is the use of subfonts. Most fonts contain glyphs for a very large number of codepoints and a pdf might only use a subset of these. To save space, a PDF creator can strip away all unneeded glyphs and create a compact subfont which will most likely use a non-standard encoding.

One workaround is to extract the font glyphs from the document, run them through OCR software and build the map from font glyph to unicode. This then lets you translate from the font-specific encoding to the unicode encoding e.g: codepoint 1 is mapped to name "c1" which, based on looking at the glyph, should be a "t", which is unicode codepoint 116.

The encoding map that you've just generated, the one going from 1 to 116, is called a ToUnicode map in the PDF standard. PDF documents can provide their own ToUnicode map, but it's optional and many do not.

## Word and paragraph detection

Reconstructing paragraphs and even words from the amorphous character soup of PDF files is a difficult task.

The PDF document provides a list of characters on a page and it is up to the consumer to identify words and paragraphs. Humans are naturally effective at

... ..



compares letter sizes, positions and alignments in order to determine what is a word/paragraph.

Naive implementations can easily have complexity larger than  $O(n^2)$ , resulting in long processing times on busy pages.

## Text and paragraph order

Deciding on text and paragraph order is difficult on two levels.

First, sometimes there is no correct answer. While documents with conventional, single column typesetting have a natural order of reading, documents with more adventurous layouts are challenging. As an example, it is not clear if the following inset should appear before, after, or during the article it is placed next to:

**Key experience:** Everything is possible in life as long as you keep fighting to reach your goal.

**Main inspiration:** My family and my first manager at SEB, Madeleine Stjernrup Öberg.

issues, implement measures and follow up on progress. SEB's Board of Directors and the Group Executive Committee adopted a governance document which states that inclusion and diversity are critical for the bank's long-term success and that SEB can and should do better in these areas.

Every year SEB conducts a Global Talent Review to identify individuals with potential for a future key role or management position.

### Labour law and unions

SEB employees are covered by collective or local agreements. SEB has a European working council with representatives from all EU and EES countries in which SEB is represented.

### Recruitment in new arenas

SEB has a strong employer brand according to annual rankings conducted among students and young professionals. This applies especially for finance and business administration students. In pace with the ongoing competence shift and growing recruitment need in new competence areas, the bank needs to strengthen its attractiveness among individuals that are attracted by IT companies and start-ups. Accordingly, SEB has widened its recruiting activities. The bank not only participates in traditional recruitment fairs for finance students but also uses interactivity and new formats such as invitations to hackathons and open workshops on artificial intelligence, blockchain technology and other cutting-edge technologies.

### SEB's core values

#### Customers first

We put our customers' needs first, always seeking to understand how to deliver real value.

#### Commitment

We are personally dedicated to the success of our customers and are accountable for our actions.

#### Collaboration

We achieve more working together.

#### Simplicity

We strive to simplify what is complex.

SEB's core values serve as the foundation for the bank's ways of working and culture, and in combination with the bank's vision – to deliver world-class service to our customers – they serve to motivate and inspire employees, managers and the organisation as a whole. These values are described in SEB's Code of Conduct, which provides guidance on ethical matters for all employees.

► Read the Code of Conduct on [sebgroup.com](https://sebgroup.com)

Second, even when the answer is clear to a human, determining robust paragraph order is a very difficult problem to solve, perhaps even AI-hard. This might sound like an extreme statement, however there are cases where the correct paragraph order can only be decided by understanding the text content.

**Prepare your utensils.**

A

Make sure you have a bowl, a chef's knife, a cutting board, spoon and fork.

**Wash the vegetables.**

C

Lukewarm water will do just fine. Make sure you do a thorough job.

**Cut the vegetables.**

B

Using the cutting board, chop all vegetables in bite-size pieces.

**Mix the vegetables and add oil.**

D

Put all the chopped vegetables in the bowl, add oil and give them all a good mix.

In the western world, a reasonable assumption is that reading is done left to right and top to bottom. So the best we can do without looking at the contents is to reduce the answer to 2 options: A B C D and A C B D.

By looking at the content, understanding what it is talking about and knowing that vegetables are washed before chopping, we can determine that A C B D is the correct order. Determining this algorithmically is a difficult problem.

That being said, a "works most times" approach is to rely on the order in which the text is stored inside the PDF document. This usually corresponds to the order the text was inserted at creation time and, for large bodies of text containing multiple paragraphs, they tend to reflect the writer-intended order.

## Embedded images

It is not uncommon for some (or all) of the PDF content to actually be a scan. In these cases, there is no text data to extract directly, so we have to resort to OCR techniques.

As an example, the Yell 2011 annual report is only available as a document scan:



2011

I'm looking for



Yell Results 2011

Search

[www.yellgroup.com/annualreport](http://www.yellgroup.com/annualreport)

## Why not OCR all the time?

While OCR might help with some of the issues shown above, it does come with its own set of drawbacks.

1. Long processing time: Running OCR on a PDF scan usually takes at least an order of magnitude longer than extracting the text directly from the PDF.
2. Difficulties with non-standard characters and glyphs: OCR algorithms have a hard time dealing with novel characters, such as smiley faces, stars/circles/squares (used in bullet point lists), superscripts, complex mathematical symbols etc.
3. No text order hints: Ordering text extracted from a PDF document is easier as the insertion order hints, most of the time, at the correct reading order. Extracting text from images offers no such hints.

## Testing

[Features](#)[Request trial](#)

extracted text is correct or expected. The approach we found most useful is to have an extensive suite of tests which look both at basic metrics (e.g. text length, page length, spaces to words ratios) and more sophisticated metrics (e.g. percentage of english words vs unrecognized words, percentage of numbers) and check for red flags such as suspicious or unexpected characters.

So what is our advice when it comes to extracting text from a PDF? Before anything else, make sure there are no better alternative data sources available.

If the data you are interested in only comes in the PDF format, then it's important to be aware that this is a deceptively simple-looking problem and that a 100% accurate solution may very well be impossible.

[← All articles](#)

© 2021 FDB Systems

FilingDB is an FDB Systems product

