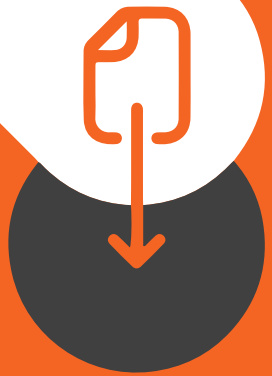
railsware



How to extract tabular data from scanned PDFs?

Proof-of-Concept made by Railsware

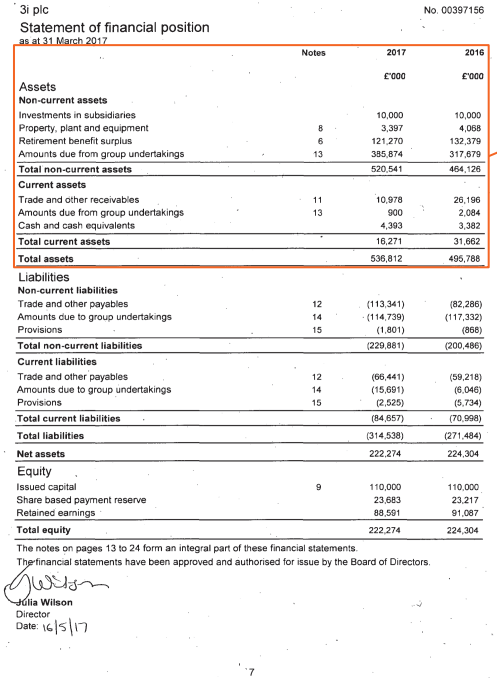
Agenda

Project overview

PDF challenges

Processing pipeline

Bonus: Heroku deployment



3i plc

No. 00397156

Statement of financial position

as at 31 March 2017

	Notes	2017	2016
		£'000	£'000
Assets			
Non-current assets			
Investments in subsidiaries		10,000	10,000
Property, plant and equipment	8	3,397	4,068
Retirement benefit surplus	6	121,270	132,379
Amounts due from group undertakings	13	385,874	317,679
Total non-current assets		520,541	464,126
Current assets			
Trade and other receivables	11	10,978	26,196
Amounts due from group undertakings	13	900	2,084
Cash and cash equivalents		4,393	3,382
Total current assets		16,271	31,662
Total assets		536,812	495,788
Liabilities			
Non-current liabilities			
Trade and other payables	12	(113,341)	(82,286)
Amounts due to group undertakings	14	(114,739)	(117,332)
Provisions	15	(1,801)	(868)
Total non-current liabilities		(229,881)	(200,486)
Current liabilities			
Trade and other payables	12	(66,441)	(59,218)
Amounts due to group undertakings	14	(15,691)	(6,046)
Provisions	15	(2,525)	(5,734)
Total current liabilities		(84,657)	(70,998)
Total liabilities		(314,538)	(271,484)
Net assets		222,274	224,304
Equity			
Issued capital	9	110,000	110,000
Share based payment reserve		23,683	23,217
Retained earnings		88,591	91,087
Total equity		222,274	224,304

The notes on pages 13 to 24 form an integral part of these financial statements.

The financial statements have been approved and authorised for issue by the Board of Directors.

Julia Wilson
Director
Date: 16/5/17

7

	Notes	2017	2016
		£'000	£'000
Assets			
Non-current assets			
Investments in subsidiaries		10,000	10,000
Property, plant and equipment	8	3,397	4,068
Retirement benefit surplus	6	121,270	132,379
Amounts due from group undertakings	13	385,874	317,679
Total non-current assets		520,541	464,126
Current assets			
Trade and other receivables	11	10,978	26,196
Amounts due from group undertakings	13	900	2,084
Cash and cash equivalents		4,393	3,382
Total current assets		16,271	31,662
Total assets		536,812	495,788

	2017	2016
Assets	£'000	£'000
Total assets	536,812	495,788

Seems simple, right?

Well… it’s not that simple…

Source: time.com

Who was the client?

Fin-tech company

UK-based

More fin than tech

Looks for automation instead of hiring people

Cares about trends on the market

What did the client want?

Ideal flow

Excel

Company number

Financial data points
extracted for last two years

Excel

What was the goal?

Semi-automated extraction

Proof of concept

3 weeks

~~27~~ 3 financial data points

~~500~~ 100 reports

What did we do?

Research & MVP

Rapid prototyping with Rails

Sidekiq-based processing pipeline

Cloud storage

3rd-party services

Let’s talk about PDF

What is a PDF?

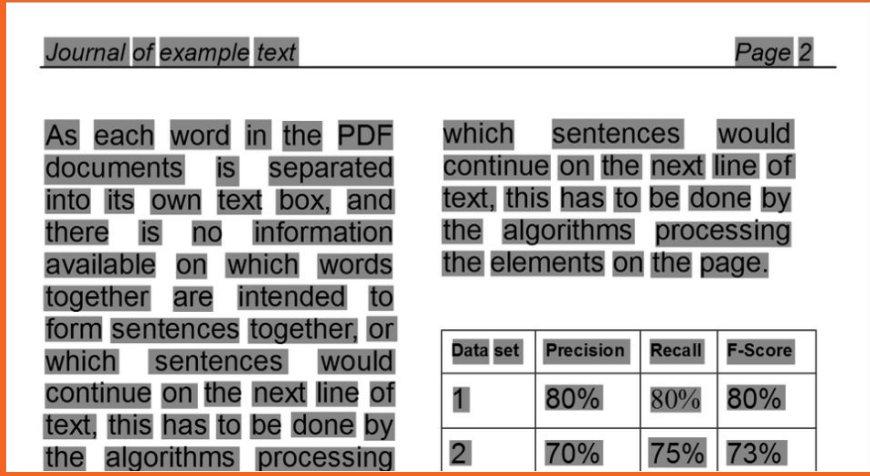
move cursor, draw, text box
PDF Reference, iText RUPS

```
((in%)) Tj
/CS0 cs
0.894 0.11 0.224 scn
/GS0 gs
/T1_2 1 Tf
6.3 0 0 6.3 135.13 690.3 Tm
[France, -3925.4, Eurozone, -2460.4,
Kingdom] TJ
15.932 1.159 Td
(United) Tj
7.173 -1.159 Td
(States) Tj
-0.206 1.159 Td
(United) Tj
```

Where are the tables then?

Tables are not there

...per se, but text elements are



Journal of example text *Page 2*

As each word in the PDF documents is separated into its own text box, and there is no information available on which words together are intended to form sentences together, or which sentences would continue on the next line of text, this has to be done by the algorithms processing the elements on the page.

Data set	Precision	Recall	F-Score
1	80%	80%	80%
2	70%	75%	73%

Source: Nurminen

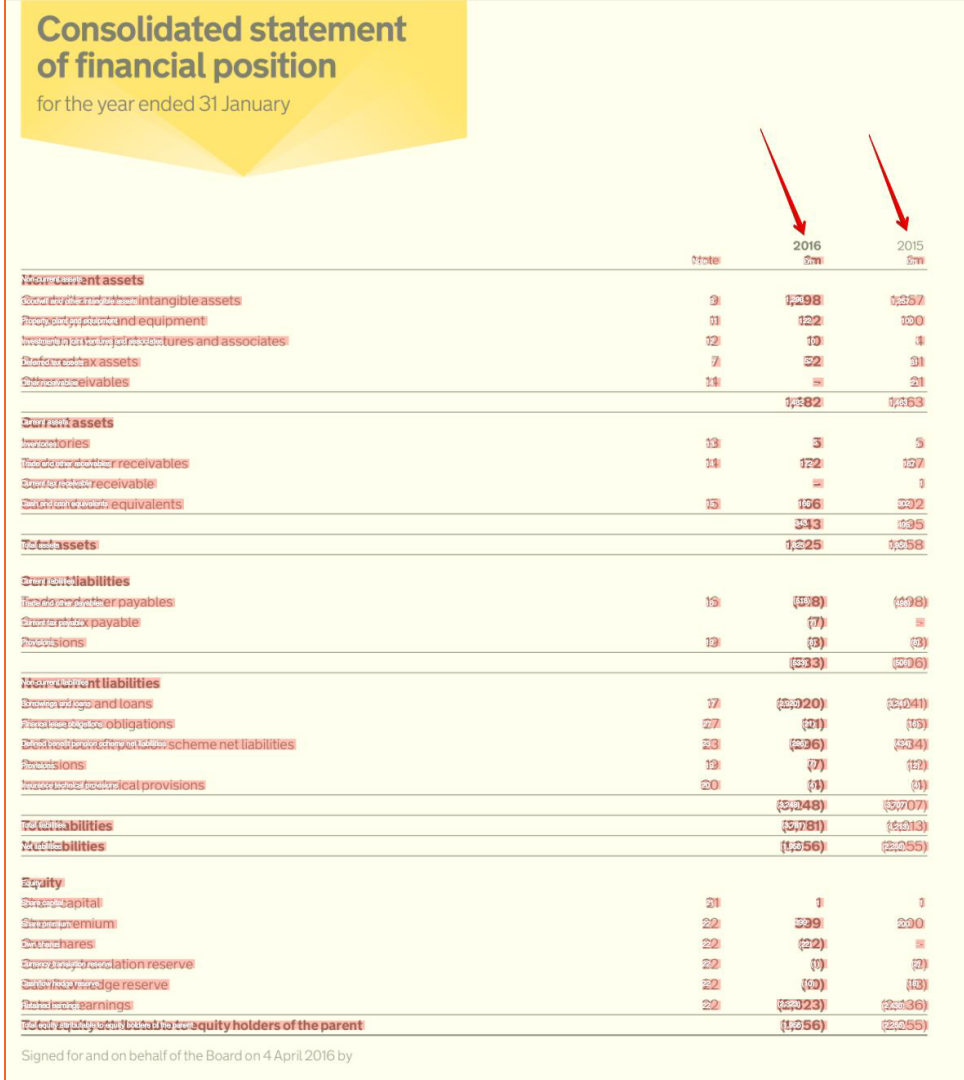
Table in PDF is not a duck - if it looks like a table, it definitely isn’t one

How do we know where a table is?

Image processing

edge detection, position analysis
Tabula, Nurminen algorithm

PDF challenges



Consolidated statement of financial position

for the year ended 31 January

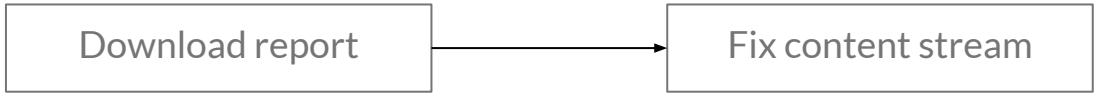
	Note	2016 £m	2015 £m
Non-current assets			
Goodwill and other intangible assets	9	1,298	1,257
Property, plant and equipment	11	122	120
Investments in joint ventures and associates	12	10	4
Deferred tax assets	7	52	31
Other receivables	14	–	21
		1,482	1,433
Current assets			
Inventories	13	3	5
Trade and other receivables	14	172	127
Derivative receivable		–	1
Cash and cash equivalents	15	136	332
		343	465
Total assets		1,825	1,898
Current liabilities			
Trade and other payables	16	(518)	(498)
Current tax payable		(7)	–
Provisions	19	(8)	(8)
		(533)	(506)
Non-current liabilities			
Borrowings and loans	17	(3,920)	(3,041)
Finance lease obligations	27	(21)	(15)
Defined benefit pension scheme net liabilities	23	(296)	(284)
Provisions	19	(7)	(12)
Insurance technical provisions	20	(1)	(1)
		(4,245)	(3,707)
Total liabilities		(4,781)	(4,213)
Net liabilities		(2,956)	(2,315)
Equity			
Share capital	21	1	1
Share premium	22	339	290
Own shares	22	(22)	–
Currency translation reserve	22	(1)	(2)
Cash flow hedge reserve	22	(60)	(43)
Retained earnings	22	(3,213)	(2,561)
Total equity attributable to equity holders of the parent		(2,956)	(2,315)

Signed for and on behalf of the Board on 4 April 2016 by

How difficult can it be?

Download report

Processing pipeline

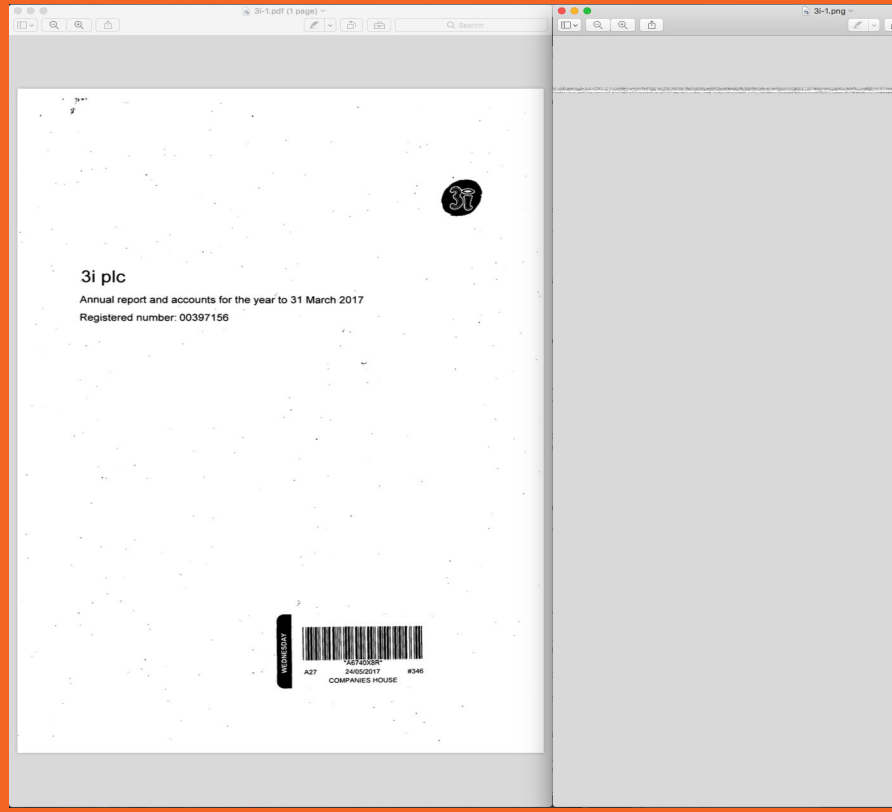


Processing pipeline

Why would you ever have to fix a PDF?

Content stream ended up unexpectedly

Output may be incorrect



How to fix the broken files?

Split pages

iText

Fast

Reliable

Quite expensive volume-based license

Maybe something else would work, though

or maybe not...

PDF Reader

Grim

Docsplit

pdftk

Apache PDF Box

How do we cut price on iText, then?

Fix PDFs before splitting

iText + Apache PDF Box

Reduce file-based license cost by 80%

Use free software to split PDFs

Processing pipeline

How do we know which pages are relevant?

We care only about balance sheets

Close to the end of a report

⅓ 40% of pages

Seems quite a lot, how can we reduce that number?

Keyword analysis of unstructured text

Google Vision API

$0.0015 per page

Every single page can be processed

~~40%~~ 8% pages get qualified

Why don’t we use Google Vision API for more?

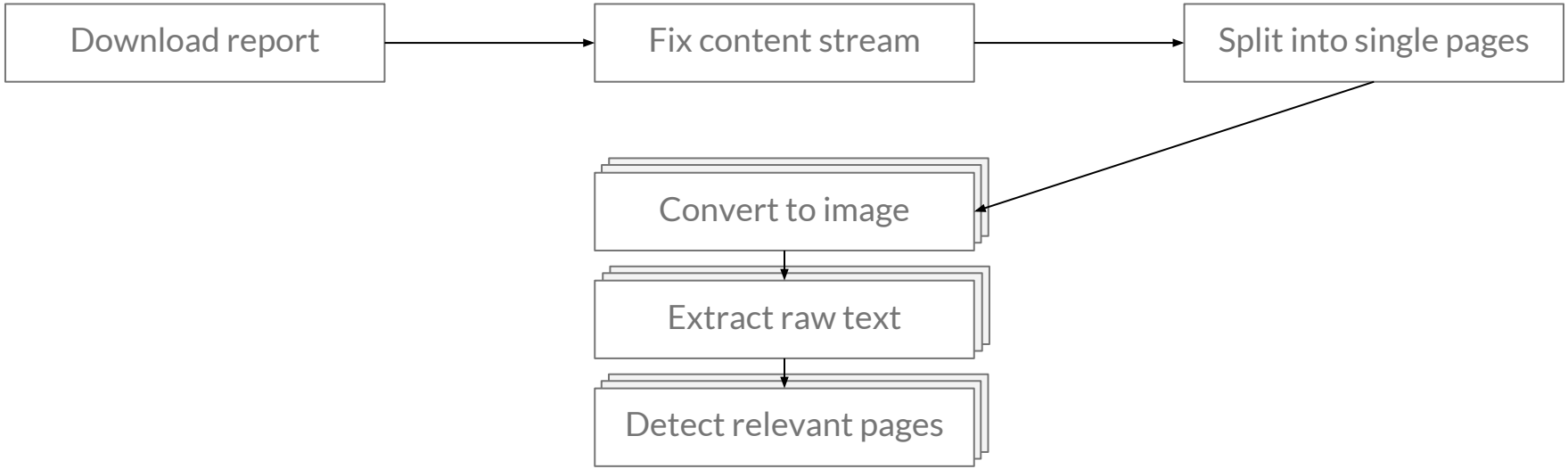
Google Vision API

OCR != layout detection

Highly specialized

OCR

Labels



Processing pipeline

OK, but how to extract tables?

Tables with text from images

Abbyy OCR SDK

OCR is never 100% correct

RTF format contains tables detected

$0.03 per page - a bit pricy...

It's cheaper to use Google Vision API to get keywords and pass only the most relevant pages to Abbyy

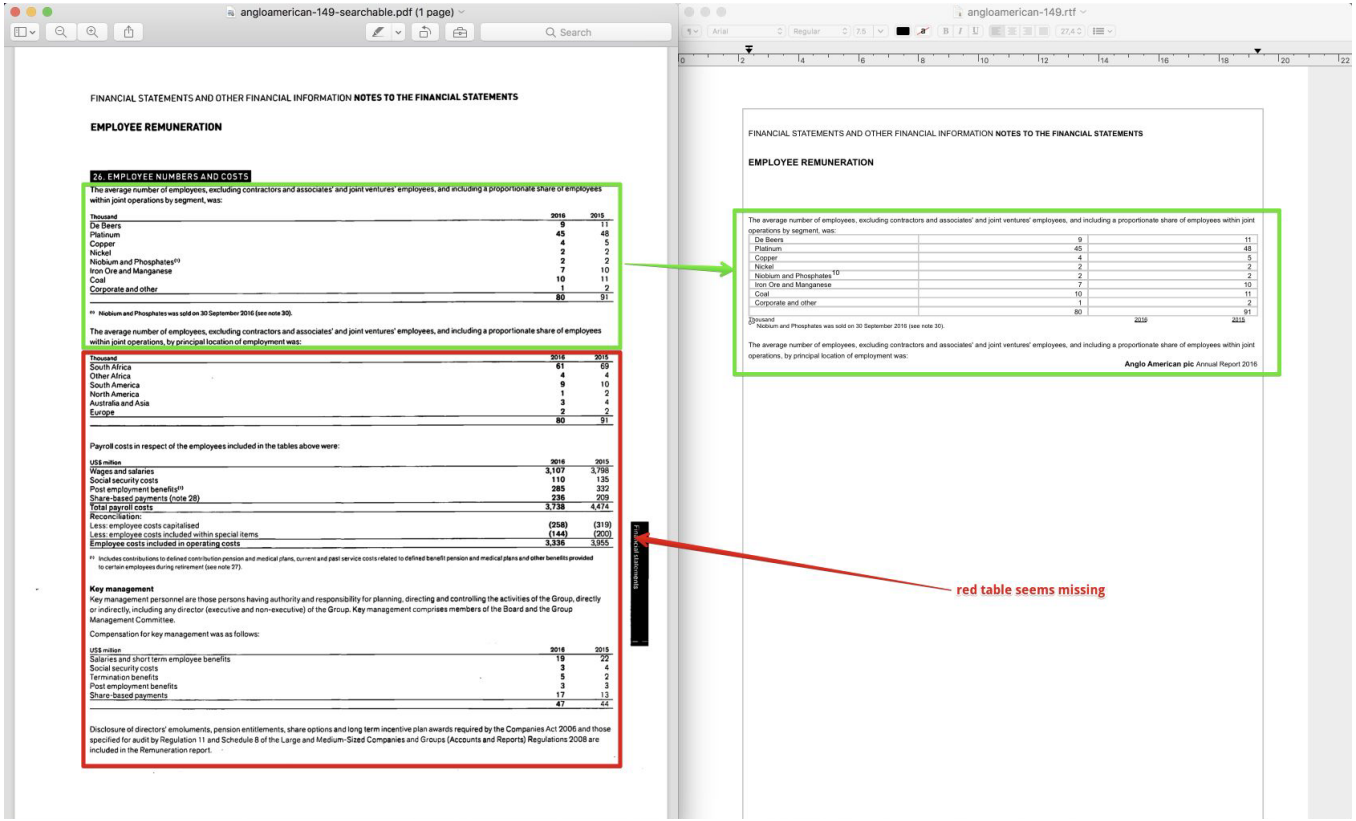
How good is Abbyy?

~30 reports assessed manually

Missing headers,
Missing tables,
text distortion

Page Id	Page link	Score	Comments
22	report_pages/22	10	
501	eport_pages/501	10	
649	eport_pages/649	10	
3812	port_pages/3812	10	unit to be taken from year
6174	port_pages/6174	10	requires table splitting
6070	port_pages/6070	10	diverted layout - years in rows, assumptions in co
1858	port_pages/1858	9	missing decimal point - just space - for 2015
327	eport_pages/327	8	years outside of table; to be restored based on co
1476	port_pages/1476	8	countries outside of table; UK and US; thankfully
2040	port_pages/2040	8	discount rate appears twice in the table; to be dis
5057	port_pages/5057	8	header row and two rows with values: past and fu
5760	port_pages/5760	8	page relevant; table detected; years outside of ta
2749	port_pages/2749	7	columns to be grouped by unit; two discount rate
2483	port_pages/2483	7	countries outside of table; UK, US, other; thankfu
4353	port_pages/4353	7	countries outside of table; UK, US, other; thankfu
5967	port_pages/5967	6	remove column with garbabe (apart from header
856	eport_pages/856	5	years outside of table; columns to be grouped by
2982	port_pages/2982	5	split countries in headers, merge countries under
1165	port_pages/1165	4	two cells (for 2016 and for 2015) combined toget
4273	port_pages/4273	3	page relevant; table detected; UK group impossib
3301	port_pages/3301	2	page relevant; table not detected
3625	port_pages/3625	2	page relevant; table not detected
5464	port_pages/5464	2	page relevant; table not detected
5504	port_pages/5504	2	page relevant; table not detected
4124	port_pages/4124	2	page relevant; table not detected
		0	where is it in the document?
		0	where is it in the document?
		0	where is it in the document?

What if tables are missing?



Tables are there… most of the time

How to read missing tables?

Convert RTF to HTML

online-convert.com

Most of the tools can handle single type of tables

Abbyy uses more than one

How to fix other problems with tables?

Pre-analysis cleanup

Try to correct common OCR errors

Insert missing headers with years

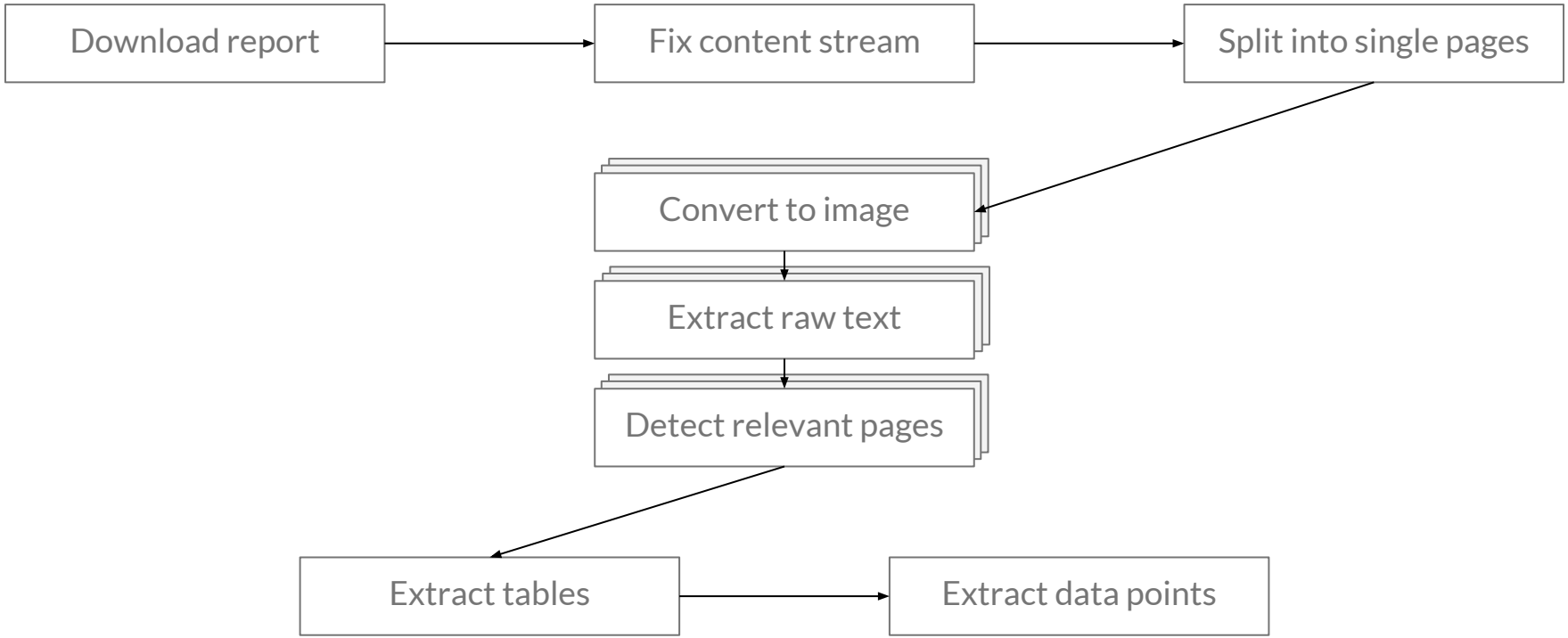
Split tables by header detection

Where is the value though?

Extraction based on a heat map

Total assets, UK, 2017

Balance sheet	2017	2017	2016	2016
£'000	UK	US	UK	US
Total assets	3,456	6,543	2,345	5,432
Total liabilities	(2,109)	(3,210)	(1,098)	(2,109)



Processing pipeline

How reliable is the extracted value?

Probability assessment

3 models tested

Weighted keywords

Value formatting and range

Other similarly good results

Wait, so it might be wrong...

Probability is calculated to bring attention to potential errors as extraction might be wrong

What's the error rate?

60%

Error rate

Meaningful data extraction from images is hard for computers

But it's easy for humans...

3i plc

No. 00397156

Statement of financial position

as at 31 March 2017

	Notes	2017	2016
		£'000	£'000
Assets			
Non-current assets			
Investments in subsidiaries		10,000	10,000
Property, plant and equipment	8	3,397	4,068
Retirement benefit surplus	6	121,270	132,379
Amounts due from group undertakings	13	385,874	317,679
Total non-current assets		520,541	464,126
Current assets			
Trade and other receivables	11	10,978	26,196
Amounts due from group undertakings	13	900	2,084
Cash and cash equivalents		4,393	3,382
Total current assets		16,271	31,662
Total assets		536,812	495,788
Liabilities			
Non-current liabilities			
Trade and other payables	12	(113,341)	(82,286)
Amounts due to group undertakings	14	(114,739)	(117,332)
Provisions	15	(1,801)	(868)
Total non-current liabilities		(229,881)	(200,486)
Current liabilities			
Trade and other payables	12	(66,441)	(59,218)
Amounts due to group undertakings	14	(15,691)	(6,046)
Provisions	15	(2,525)	(5,734)

How to automate this?

Mechanical Turk

Scalable workforce from Amazon

Comparable cost per page to Abbyy

Pay only for approved answers

Automated cross-check policy

How does it work?

amazon mturk Worker | HITs | Dashboard

Skip | Accept

HIT Details

Reward: $0.01

Time Allotted: 00:10:00

You must accept this Requester's HIT before working on it. Learn more

iframe with HTML question or external question

Type the text, very carefully.

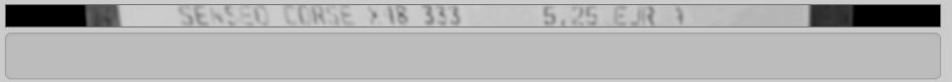
Warning

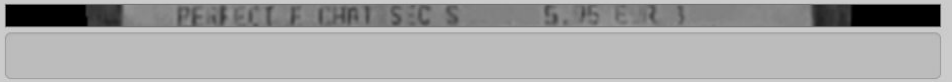
Your HIT won't be approved if you don't stick to the following instructions

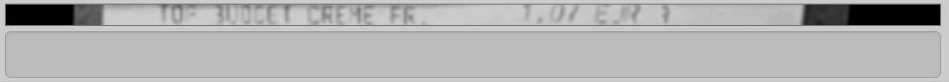
Instructions

- If a **€** symbol is present and not on your keyboard, **copy/paste** it from this line: €
- Other **special characters** that you may need: **% & * ° < >**
- Be careful, "ABCD" and "abcd" are **different**.
- If you cannot read a text. **NEVER EVER** write "I can't read". If you do write it, your HIT will be rejected and we may block you.
  Two options:
  - A checkbox **I cannot read this text** is available: justtick it!
  - The checkbox is not available: leave the text input blank, it won't be penalized.
- When there are 1, 2 or more spaces, type just **1** space.
- If no image is loading, **do NOT accept this HIT**.

More instructions here







Submit

Report this HIT | Why Report

Skip | Accept

What's the error rate?

60%

Error rate for
Automated extraction

18%

Error rate for
Mechanical Turk

Where do we go from here?

Next steps

RW Labs

More reliable probability model

Limit relevant pages

Add connector for Mechanical Turk

...

AI?

What have we learned?

Lessons learned

~~Build~~ ~~measure~~ learn

Automation is a part of the picture

Human supervision is necessary

Mechanical Turk - valid alternative

Start simple and be pragmatic

Cost estimation is important for development

BONUS: How to configure Heroku for all of that?

Buildpacks

[Buildpack API]

Ruby

Node.js

Java

Google Cloud SDK

Why do you need Java?

java -jar ...

iText, Apache PDF Box

Requires pom.xml

```
<project NAMESPACES_HERE>
<modelVersion>4.0.0</modelVersion>
<groupId>GROUP_HERE</groupId>
<artifactId>ARTIFACT_HERE</artifactId>
<version>VERSION_HERE</version>
</project>
```

And some memory...

What about Google Cloud SDK?

Google Cloud SDK

heroku-google-cloud-buildpack

Download and install SDK

Build JSON with credentials from ENV variables

Activate service account using profile.d script

Questions?

Thank you