EXTRACTING TABLE DATA FROM PDF

A JOURNEY INTO THE UNKNOWN

by Don Vito Perleone

German Perl Workshop 2016
Nürnberg

BASED ON A TRUE STORY

ABOUT ME

- 20 years of commercial Perl experience
- Enjoy messing about with data
- Hannover.pm

DR. STRANGELOVE – WHY?

REAL-WORLD EXAMPLE

- Cafeteria at $work, run by external service provider
- Weekly menu available as PDF

Speiseplan KW 05

	Montag	Dienstag	Mittwoch	Donnerstag	Freitag
	1. Februar 2016	2. Februar 2016	3. Februar 2016	4. Februar 2016	5. Februar 2016
Suppe 0,85 €	Tomatencremesuppe	Gemüsebrühe[28]	Kartoffelsuppe[28]	Zwiebelsuppe	Erbsencremsuppe[26,28]
Hauptgerichte	**Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis** 3,60 €	**Riesencurrywurst[2,3,8] mit zweierlei Saucen[1,2] und Pommes Frites** 3,60 €	**Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis[27] und Grillgemüse[28]** 3,60 €	**Gebackener Leberkäse[20w] mit Bohnengemüse[26] und Kartoffelstampf** 3,60 €	**Schlemmerfilet á la Bordelaise[4,20] mit Dillschaum[20,26] Tomatensalat und Kartoffeln** 3,60 €
	Putengulasch mit Zucchini und Champignon[20w,26,28] dazu Nudeln[20w,23] 3,60 €	**Kutterfrikadelle[4,21] Gurkensalat[2,3] und Kartoffelpüree** 3,60 €	**Hausgemachtes Schnitzel[20w] "Mailänder Art" mit Spaghetti[20w] und Tomatensauce** 3,60 €	**Highlight der Woche** **Hirschrückensteak an Rotweinjus, Rosenkohl und Butterspätzle[20w,]** 5,90 €	**Tanduri Masala Putenbraten auf Gemüse-Cous-Cous mit Joghurtdip[26]** 3,60 €
Vegetarisch	**Fit Menü** **Gratinierte Maultaschen [20(W),23,28] auf Tomatenspinat** 3,30 €	**Fit Menü** **Frühlingsrolle[2,4,20,25,30] mit Süß-Saurem Dip und Basmatireis** 3,30 €	**Fit Menü** **Gemüselasagne[20w,26,28] mit Basilikum - Tomatensauce** 3,30 €	**Fit Menü** **Maisplinsen[20w,23,26] an gegrilltem Gemüse mit Limettensauce[26]** 3,30 €	**Fit Menü** **Rahmige Kartoffelsuppe[26,28] mit bunter Gemüseeinlage und Baguette[20,23,25,26,28,29,30,32]** 3,30 €
Salate	**Bei der Zubereitung der Speisen verwenden wir Allergie / Lebensmittelunverträglichkeit auslösende Stoffe als Zutaten, diese sind im Speiseplan gekennzeichnet. Trotz größtmöglicher Sorgfalt können wir jedoch nicht ausschließen, dass technisch unvermeidbare Einträge von Allergie / Lebensmittelunverträglichkeit auslösenden Stoffen und ggf. weitere mögliche Einträge auch in andere Gerichte gelangen. Weitere Auskünfte dazu erteilt Ihnen gerne unser Küchenpersonal.**				
Beilagen 0,70 €	Wildreis Nudeln Tagesgemüse[28]	Pommes Frites Kartoffelpüree Basmatireis Tagesgemüse[28]	Nussreis Spaghetti Tagesgemüse[28]	Kartoffelstampf Spätzle Tagesgemüse[28]	Kartoffelsalat Reis Tagesgemüse[28]
Dessert Groß 0,80 € Klein 0,50 €					

Zusatzstoffe sind mit Zahlen gekennzeichnet siehe Aushang
Bei Fragen zu Dussmann können Sie auf www.dussmann.com klicken!

Speiseplanänderungen vorbehalten

Speiseplan KW 06

	Montag	Dienstag	Mittwoch	Donnerstag	Freitag
	8. Februar 2016	9. Februar 2016	10. Februar 2016	11. Februar 2016	12. Februar 2016
Suppe 0,85 €	Hühnerbrühe mit Einlage	Käse - Lauchsuppe	Minestrone	Karottencremesuppe mit Ingwer[26,28]	Pfifferlingcremesuppe
Hauptgerichte	**Gefüllte Paprikaschote mit Schweinemett, Reis und Tomatensauce** **3,60 €**	**Crispy Chicken Burger mit Asiasauce und Pommes Frites[20,23,26]** **3,60 €**	**Kohlrouladen mit Hackfleischfüllung,[20w,26,29] Kohlgemüse und Salzkartoffeln** **3,60 €**	**Griechische Bratkartoffeln mit Hähnchenbruststreifen Zucchini und Hirtenkäse[26]** **3,60 €**	**Grünkohl[20H,28,29] mit Kassler, Mettende[1,2,3] und Kartoffeln** **3,80 €**
	Ente Süß-Sauer[1,2,3,9,20w,25,28,29] mit gebratenen Nudeln **3,60 €**	**Gedünsteter Tilapia[21] "Lingurische Art" mit Oliven und Tomaten[28] dazu Pasta[20w,23]** **3,60 €**	**Grüne Thaisuppe[20] mit Hähnchenbrust und Reisnudeln** **3,60 €**	**Highlight der Woche** **Gebratene Forelle auf Gemüse- Kartoffel-Bett mit Tomaten - Mozarellasauce** **5,20 €**	**Deftiger Bohneneintopf mit Kassler und Baguette[20w,26,28]** **3,60 €**
Vegetarisch	**Fit Menü** **Blumenkohl "Indisch" mit Nüssen, Curry, Rosinen[24,26,27] und Basmatireis** **3,30 €**	**Fit Menü** **Milchreis[26] mit Heißen Kirschen und Zimt Zucker** **3,30 €**	**Fit Menü** **Penne[20] Napoli und Grana Padano[23,26]** **3,30 €**	**Fit Menü** **Ofenkartoffel mit zweierlei Quark[26] dazu Salatbeilage** **3,30 €**	**Fit Menü** **Gekochte Eier[23] in Dijonsenfrahm[20w,28,29] mit Salzkartoffeln und Rote Beete Salat[9]** **3,30 €**
Salate	**Bei der Zubereitung der Speisen verwenden wir Allergie / Lebensmittelunverträglichkeit auslösende Stoffe als Zutaten, diese sind im Speiseplan gekennzeichnet. Trotz größtmöglicher Sorgfalt können wir jedoch nicht ausschließen, dass technisch unvermeidbare Einträge von Allergie / Lebensmittelunverträglichkeit auslösenden Stoffen und ggf. weitere mögliche Einträge auch in andere Gerichte gelangen. Weitere Auskünfte dazu erteilt Ihnen gerne unser Küchenpersonal.**				
Beilagen 0,70 €	Reis gebratene Nudeln Basmatireis Tagesgemüse[28]	Pommes Frites Pasta Tagesgemüse[28]	Kartoffeln Reisnudeln Penne Tagesgemüse[28]	Griechische Bratkartoffeln Ofenkartoffel Tagesgemüse[28]	Kartoffeln Reis Tagesgemüse[28]
Dessert Groß 0,80 € Klein 0,50 €					

Zusatzstoffe sind mit Zahlen gekennzeichnet siehe Aushang
Bei Fragen zu Dussmann können Sie auf www.dussmann.com klicken!

Speiseplanänderungen vorbehalten

Speiseplan KW 05

	Montag	Dienstag	Mittwoch
	1. Februar 2016	2. Februar 2016	3. Februar 2016
Suppe 0,85 €	Tomatencremesuppe	Gemüsebrühe[28]	Kartoffelsuppe[28]
Haupt-gerichte	Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis 3,60 €	Riesencurrywurst[2,3,8] mit zweierlei Saucen[1,2] und Pommes Frites 3,60 €	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis[27] und Grillgemüse[28] 3,60 €
	Putengulasch mit Zucchini und Champignon[20w,26,28] dazu Nudeln[20w,23] 3,60 €	Kutterfrikadelle[4,21] Gurkensalat[2,3] und Kartoffelpüree 3,60 €	Hausgemachtes Schnitzel[20w] "Mailänder Art" mit Spaghetti[20w] und Tomatensauce 3,60 €
Vegeta-	Fit Menü	Fit Menü	Fit Menü
	Gratinierte Maultaschen	Frühlingsrolle[2,4,20,25,30]	Gemüselasagne[20w,26,28] mit

Speiseplan KW 06

	Montag	Dienstag	Mittwoch
	8. Februar 2016	9. Februar 2016	10. Februar 2016
Suppe 0,85 €	Hühnerbrühe mit Einlage	Käse - Lauchsuppe	Minestrone
Haupt-gerichte	Gefüllte Paprikaschote mit Schweinemett, Reis und Tomatensauce 3,60 €	Crispy Chicken Burger mit Asiasauce und Pommes Frites[20,23,26] 3,60 €	Kohlrouladen mit Hackfleischfüllung,[20w,26,29] Kohlgemüse und Salzkartoffeln 3,60 €
	Ente Süß-Sauer[1,2,3,9,20w,25,28,29] mit gebratenen Nudeln 3,60 €	Gedünsteter Tilapia[21] "Lingurische Art" mit Oliven und Tomaten[28] dazu Pasta[20w,23] 3,60 €	Grüne Thaisuppe[20] mit Hähnchenbrust und Reisnudeln 3,60 €
Vegeta-	Fit Menü Blumenkohl "Indisch" mit Nüssen, Curry,	Fit Menü Milchreis[26] mit Heißen	Fit Menü Penne[20] Napoli und Grana

FOOD RATING APPLICATION "GOURMETER"

- Made by our software engineering trainee
- Needs weekly manual input of the menus
- Luckily not by me
- But I do enjoy scripting things
- and making them run on their own

AUTOMATION BECKONS

- Not as easy as it looks
- Naive approach doesn't work
- ... at all
- Dig deeper
- Much deeper

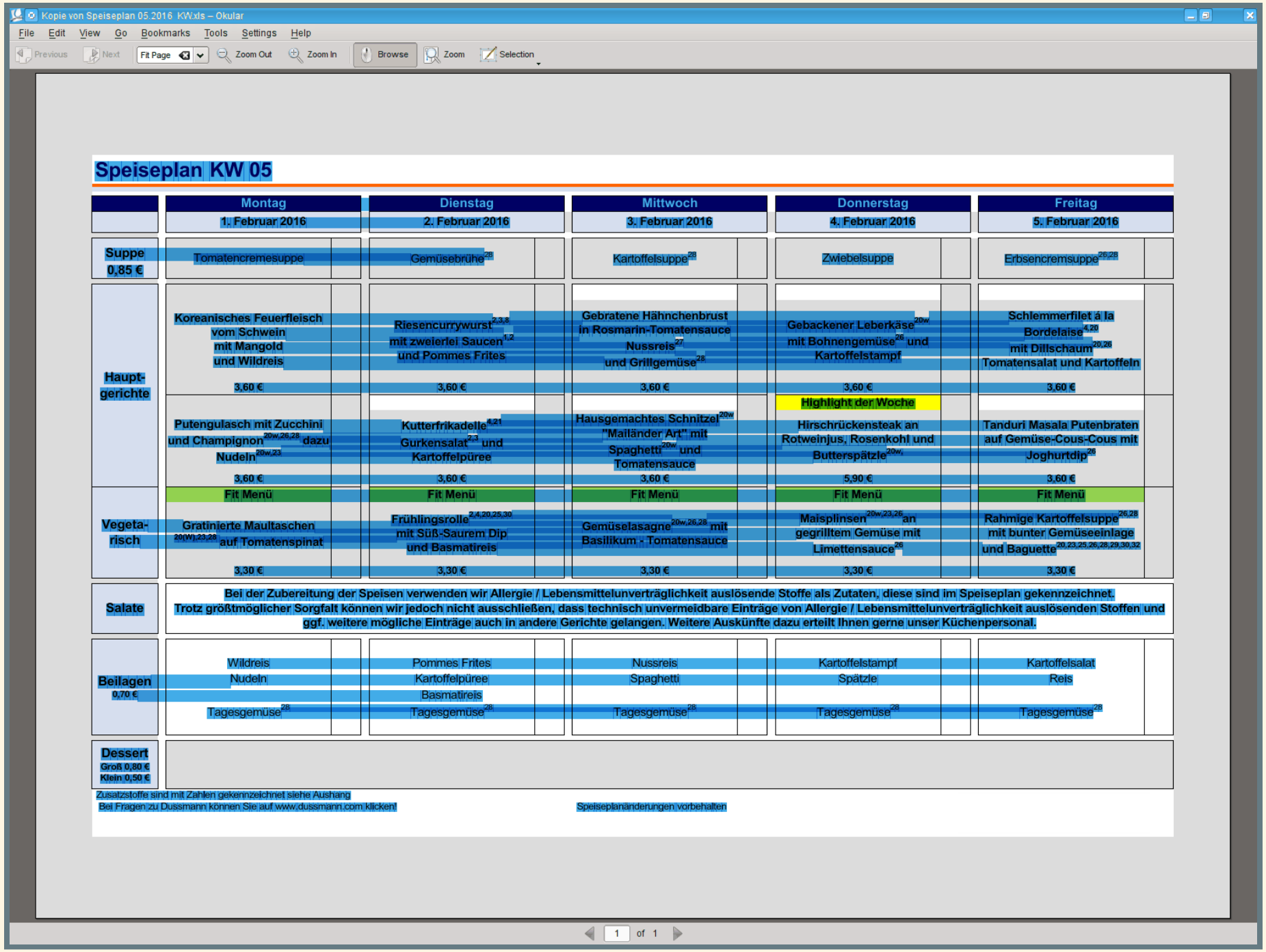
WHY DID I DO THIS?

- London Perl Workshop last December
- Attended talk "Perl and PDF" by Prabhakar Somu
- `http://act.yapc.eu/lpw2015/talk/6423`
- `https://youtu.be/5_yFwl5zN8w`
- Motivated me to submit, research and prepare talk

FIRST BLOOD – NAIVE APPROACH

COPYING THE TEXT

- It's right there, I can see it
- I can also select it



Speiseplan KW 05

	Montag 1. Februar 2016	Dienstag 2. Februar 2016	Mittwoch 3. Februar 2016	Donnerstag 4. Februar 2016	Freitag 5. Februar 2016
Suppe **0,85 €**	Tomatencremesuppe	Gemüsebrühe[28]	Kartoffelsuppe[28]	Zwiebelsuppe	Erbsencremsuppe[26,28]
Haupt-gerichte	**Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis** **3,60 €**	**Riesencurrywurst[2,3,8] mit zweierlei Saucen[1,2] und Pommes Frites** **3,60 €**	**Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis[27] und Grillgemüse[28]** **3,60 €**	**Gebackener Leberkäse[20w] mit Bohnengemüse[26] und Kartoffelstampf** **3,60 €**	**Schlemmerfilet á la Bordelaise[4,20] mit Dillschaum[20,26] Tomatensalat und Kartoffeln** **3,60 €**
	Putengulasch mit Zucchini und Champignon[20w,26,28] dazu Nudeln[20w,23] **3,60 €**	**Kutterfrikadelle[4,21] Gurkensalat[2,3] und Kartoffelpüree** **3,60 €**	**Hausgemachtes Schnitzel[20w] "Mailänder Art" mit Spaghetti[20w] und Tomatensauce** **3,60 €**	**Highlight der Woche** **Hirschrückensteak an Rotweinjus, Rosenkohl und Butterspätzle[20w,]** **5,90 €**	**Tanduri Masala Putenbraten auf Gemüse-Cous-Cous mit Joghurtdip[26]** **3,60 €**
Vegeta-risch	**Fit Menü** **Gratinierte Maultaschen[20(W),23,28] auf Tomatenspinat** **3,30 €**	**Fit Menü** **Frühlingsrolle[2,4,20,25,30] mit Süß-Saurem Dip und Basmatireis** **3,30 €**	**Fit Menü** **Gemüselasagne[20w,26,28] mit Basilikum - Tomatensauce** **3,30 €**	**Fit Menü** **Maisplinsen[20w,23,26] an gegrilltem Gemüse mit Limettensauce[26]** **3,30 €**	**Fit Menü** **Rahmige Kartoffelsuppe[26,28] mit bunter Gemüseeinlage und Baguette[20,23,25,26,28,29,30,32]** **3,30 €**
Salate	**Bei der Zubereitung der Speisen verwenden wir Allergie / Lebensmittelunverträglichkeit auslösende Stoffe als Zutaten, diese sind im Speiseplan gekennzeichnet. Trotz größtmöglicher Sorgfalt können wir jedoch nicht ausschließen, dass technisch unvermeidbare Einträge von Allergie / Lebensmittelunverträglichkeit auslösenden Stoffen und ggf. weitere mögliche Einträge auch in andere Gerichte gelangen. Weitere Auskünfte dazu erteilt Ihnen gerne unser Küchenpersonal.**				
Beilagen **0,70 €**	Wildreis Nudeln Tagesgemüse[28]	Pommes Frites Kartoffelpüree Basmatireis Tagesgemüse[28]	Nussreis Spaghetti Tagesgemüse[28]	Kartoffelstampf Spätzle Tagesgemüse[28]	Kartoffelsalat Reis Tagesgemüse[28]
Dessert **Groß 0,80 €** **Klein 0,50 €**					

Zusatzstoffe sind mit Zahlen gekennzeichnet siehe Aushang
Bei Fragen zu Dussmann können Sie auf www.dussmann.com klicken!

Speiseplanänderungen vorbehalten

COPY TEXT

- It's right there, I can see it
- I can also select it
- And copy it
- But the result is
- not quite what you expect

```
Speiseplan KW 05
Montag
 Dienstag
1. Februar 2016
 2. Februar 2016
Suppe
 28
Tomatencremesuppe
 Gemüsebrühe
0,85 €
Mittwoch
3. Februar 2016
Kartoffelsuppe
28
Donnerstag
4. Februar 2016
Zwiebelsuppe
Freitag
5. Februar 2016
Erbsencremsuppe
26,28
Koreanisches Feuerfleisch
 2,3,8
 Gebratene Hähnchenbrust
```

```
24    Gebratene Hähnchenbrust
25    20w
26    Schlemmerfilet á la
27  Riesencurrywurst Gebackener Leberkäse 4,20
28  vom Schwein
```

COMMAND LINE TEXT EXTRACTION

- Several tools
  - `ps2txt` (ghostscript)
  - `pstotxt` (also ghostscript)
  - `pdftotext` (xpdf, poppler)
- Output ranges from total crap
- to all there, but wrong order

```

\Delta \Theta \Lambda \Lambda \Xi \Pi \Sigma \Upsilon \Phi \Psi \Omega

fffifl\Xi ffifflij`'\Xi j\Psi ^\Xi \Theta \Xi fl*_\Xi ij`'*fi,\Psi \Delta `'ss

\Theta fflAE\Psi OEi_AEfl\Xi ij

O/i\Xi j\Xi ffl`\Theta flflss\Theta fljae!\Sigma "\Sigma \Upsilon ,iae\Psi #ss

\Theta fflAE\Psi &fli__o/\Xi ,.j\Xi !\Upsilon

&\Xi 'ffi`/\Xi ffl\Xi fl\Psi 0\Xi '\Xi fl/)j\Xi !\Pi ss\Psi ,iae\Psi 1fi'ffl\X
```

```
 1 12334
 2 56789\Omega
 3 fffifl4ffiifflij``4j9<9F>424fl<96>4ij`` <9A>fi91``4iffl iae9oeffiifflfiAE 2fflAE
 4 i4j4ffl`2flfl/2fljae !6''67
 5 iae9#4i4fl4i91ffi2`4ffl $6!
 6 2fflAE9%fi4j9<9F>fliae4j
 7 &4'flffiae4ffl49()'ffl``4ffl'fl2jae
 8 iffl9fijffifliffl*+fiffiae4ffljffi2`4 ,2jjfl4ij !
 9 2fflAE9&fli4.j4 !7
10 &4'ffi`/4ffl4fl904'4fl/)j4 !59
11 iae91fi'ffl4ffl4.j4 !2 92fflAE9
12 ffffiflaefi<96><96>4jaeffi3<96>
13 1``44fl<96>i4ae939ffi9 1fiflAE4ffiij4 46!5
14 iae95ij``ffi2 !56!29
15 +fiffiae4ffljffiffiae92fflAE9ffffiflaefi<96><96>4ffl
16 ''6259\Omega ''6259\Omega ''6259\Omega ''6259\Omega ''6259\Omega (i'i'ae9AE4fl
17 %2ae4ffl2ffij``9iae962``iffli9
18 2fflAE97'ffi3ifflfiffl !56!26!7 9AEffi#29
19 ,2AE4ffl !56!''
20 ff2aeae4fl<96>fli/ffiAE44 46!$9
21 &2fl/4ffljffiffiae !6'' 92fflAE9
22 ffffiflaefi<96><96>43.fl44
23 (ffi2j4ffi``ae4j91``ffliae#4 !59
24 8oeffii)fflAE4fl99flae89iae9 13ffi'4aeaei !5 92fflAE9
```

Speiseplan KW 05
Montag

Suppe
0,85 €

Dienstag

Mittwoch

Donnerstag

Freitag

1. Februar 2016

2. Februar 2016

3. Februar 2016

4. Februar 2016

5. Februar 2016

Tomatencremesuppe

Koreanisches Feuerfleisch
vom Schwein
mit Mangold
und Wildreis

Hauptgerichte

Riesencurrywurst2,3,8
mit zweierlei Saucen1,2
und Pommes Frites

Gebratene Hähnchenbrust
in Rosmarin-Tomatensauce
Nussreis27
und Grillgemüse28

3,60 €

3,60 €

3,60 €

Zwiebelsuppe

Erbsencremsuppe

QUICK SIDEBAR

HOW I BUILT THE PREVIOUS TWO-COLUMN SLIDE

```
$ paste -d=
<( cat -n menu.txt | sed -n 1,25p )
<( cat -n menu.txt | sed -n 35,+24p )
| column -s= -t | expand
```

PDF ...

- Short for Portable **Document** Format
- Concerned with **Looks** and **Layout**
- Structure?
- **What** structure?

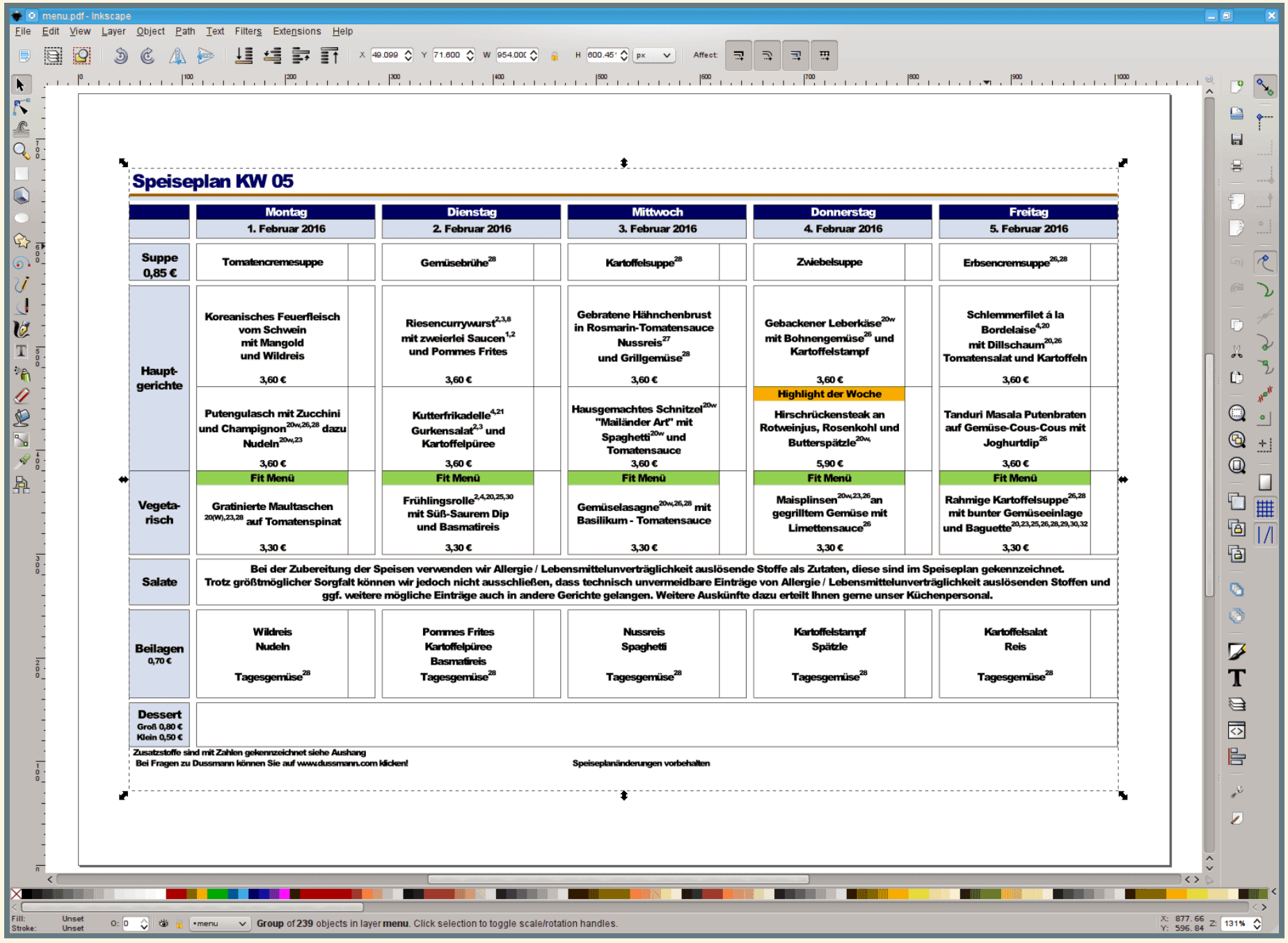
SAVING PRIVATE RYAN – SNEAKY APPROACH

WE'RE GONNA NEED A BIGGER CALIBER!

- SVG
- Libre Office
- Last Resort: OCR

SVG

- Text based vector format
- Inkscape
- ImageMagick
- pdftocairo (poppler)
- Grouping not useful



Speiseplan KW 05

	Montag 1. Februar 2016	Dienstag 2. Februar 2016	Mittwoch 3. Februar 2016	Donnerstag 4. Februar 2016	Freitag 5. Februar 2016
Suppe 0,85 €	Tomatencremesuppe	Gemüsebrühe[28]	Kartoffelsuppe[28]	Zwiebelsuppe	Erbsencremsuppe[26,28]
Haupt-gerichte	Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis 3,60 €	Riesencurrywurst[2,3,8] mit zweierlei Saucen[1,2] und Pommes Frites 3,60 €	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis[27] und Grillgemüse[28] 3,60 €	Gebackener Leberkäse[20w] mit Bohnengemüse[26] und Kartoffelstampf 3,60 €	Schlemmerfilet á la Bordelaise[4,20] mit Dillschaum[20,26] Tomatensalat und Kartoffeln 3,60 €
	Putengulasch mit Zucchini und Champignon[20w,26,28] dazu Nudeln[20w,23] 3,60 €	Kutterfrikadelle[4,21] Gurkensalat[2,3] und Kartoffelpüree 3,60 €	Hausgemachtes Schnitzel[20w] "Mailänder Art" mit Spaghetti[20w] und Tomatensauce 3,60 €	**Highlight der Woche** Hirschrückensteak an Rotweinjus, Rosenkohl und Butterspätzle[20w,] 5,90 €	Tanduri Masala Putenbraten auf Gemüse-Cous-Cous mit Joghurtdip[26] 3,60 €
Vegeta-risch	**Fit Menü** Gratinierte Maultaschen[20(W),23,28] auf Tomatenspinat 3,30 €	**Fit Menü** Frühlingsrolle[2,4,20,25,30] mit Süß-Saurem Dip und Basmatireis 3,30 €	**Fit Menü** Gemüselasagne[20w,26,28] mit Basilikum - Tomatensauce 3,30 €	**Fit Menü** Maisplinsen[20w,23,26] an gegrilltem Gemüse mit Limettensauce[26] 3,30 €	**Fit Menü** Rahmige Kartoffelsuppe[26,28] mit bunter Gemüseeinlage und Baguette[20,23,25,26,28,29,30,32] 3,30 €
Salate	Bei der Zubereitung der Speisen verwenden wir Allergie / Lebensmittelunverträglichkeit auslösende Stoffe als Zutaten, diese sind im Speiseplan gekennzeichnet. Trotz größtmöglicher Sorgfalt können wir jedoch nicht ausschließen, dass technisch unvermeidbare Einträge von Allergie / Lebensmittelunverträglichkeit auslösenden Stoffen und ggf. weitere mögliche Einträge auch in andere Gerichte gelangen. Weitere Auskünfte dazu erteilt Ihnen gerne unser Küchenpersonal.				
Beilagen 0,70 €	Wildreis Nudeln Tagesgemüse[28]	Pommes Frites Kartoffelpüree Basmatireis Tagesgemüse[28]	Nussreis Spaghetti Tagesgemüse[28]	Kartoffelstampf Spätzle Tagesgemüse[28]	Kartoffelsalat Reis Tagesgemüse[28]
Dessert Groß 0,80 € Klein 0,50 €					

Zusatzstoffe sind mit Zahlen gekennzeichnet siehe Aushang
Bei Fragen zu Dussmann können Sie auf www.dussmann.com klicken!

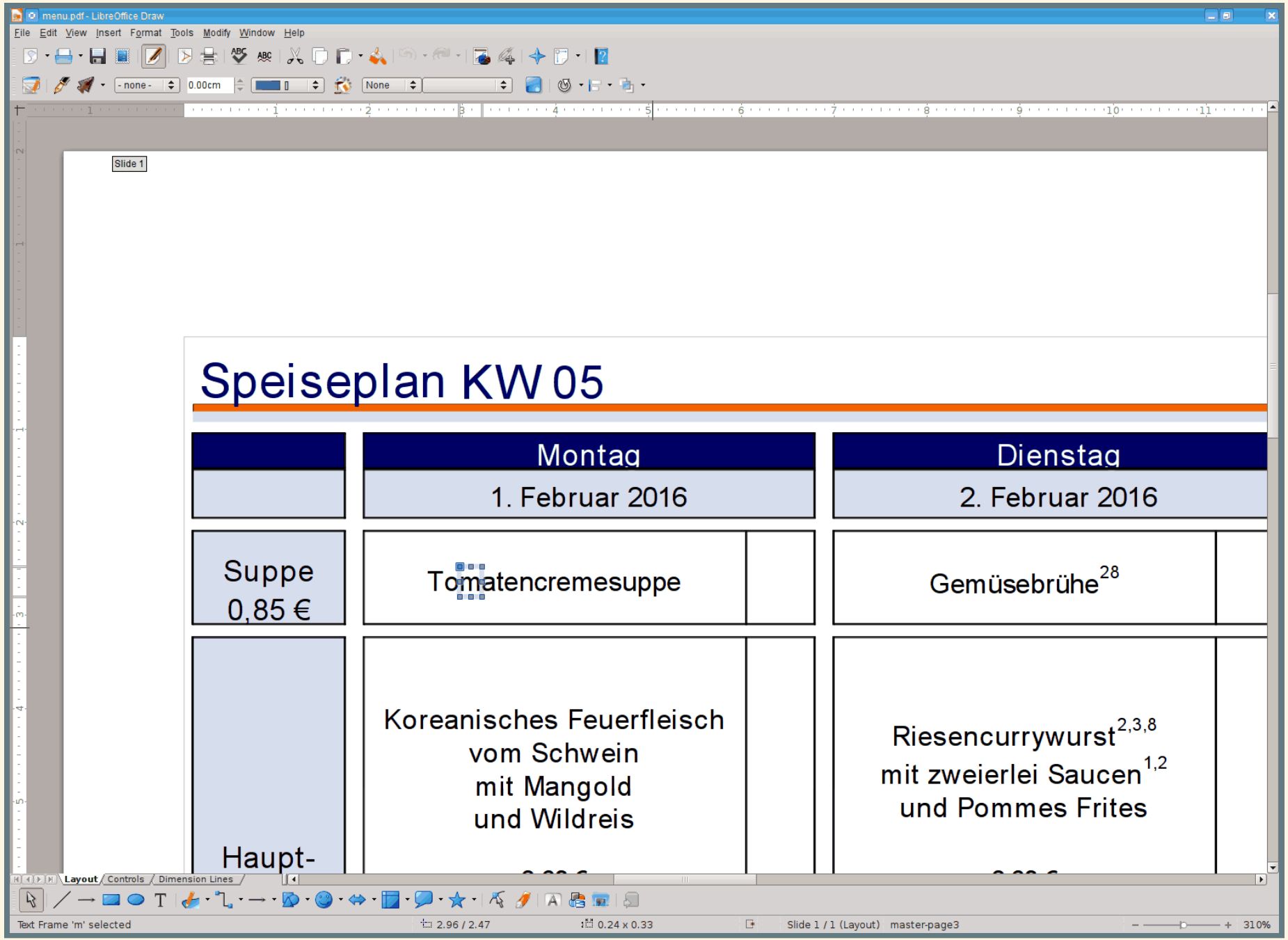
Speiseplanänderungen vorbehalten

Speiseplan KW 05

	Montag	Dienstag	Mittwoch	Don
	1. Februar 2016	2. Februar 2016	3. Februar 2016	4. Feb
Suppe 0,85 €	Tomatencremesuppe	Gemüsebrühe[28]	Kartoffelsuppe[28]	Zwiebe
Haupt-gerichte	Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis 3,60 €	Riesencurrywurst[2,3,8] mit zweierlei Saucen[1,2] und Pommes Frites 3,60 €	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis[27] und Grillgemüse[28] 3,60 €	Gebackener I mit Bohneng Kartoffe 3,6
	Putengulasch mit Zucchini und Champignon[20w,26,28] dazu Nudeln[20w,23] 3,60 €	Kutterfrikadelle[4,21] Gurkensalat[2,3] und Kartoffelpüree 3,60 €	Hausgemachtes Schnitzel[20w] "Mailänder Art" mit Spaghetti[20w] und Tomatensauce 3,60 €	Highlight d Hirschrück Rotweinjus, R Buttersp 5,9
Vegeta-risch	Fit Menü Gratinierte Maultaschen[20(W),23,28] auf Tomatenspinat 3,30 €	Fit Menü Frühlingsrolle[2,4,20,25,30] mit Süß-Saurem Dip und Basmatireis 3,30 €	Fit Menü Gemüselasagne[20w,26,28] mit Basilikum - Tomatensauce 3,30 €	Fit M Maisplinse gegrilltem G Limetten 3,3

LIBRE OFFICE

- Can open PDF files
- Imported as drawing
- Each letter a separate element
- Not suited for table extraction


menu.pdf - LibreOffice Draw
File Edit View Insert Format Tools Modify Window Help
Slide 1
Speiseplan KW 05
Montag
1. Februar 2016
Dienstag
2. Februar 2016
Suppe
0,85 €
Tomatencremesuppe
Gemüsebrühe28
Haupt-
Koreanisches Feuerfleisch
vom Schwein
mit Mangold
und Wildreis
Riesencurrywurst2,3,8
mit zweierlei Saucen1,2
und Pommes Frites
Layout
Controls
Dimension Lines
Text Frame 'm' selected
2.96 / 2.47
0.24 x 0.33
Slide 1 / 1 (Layout)
master-page3
310%

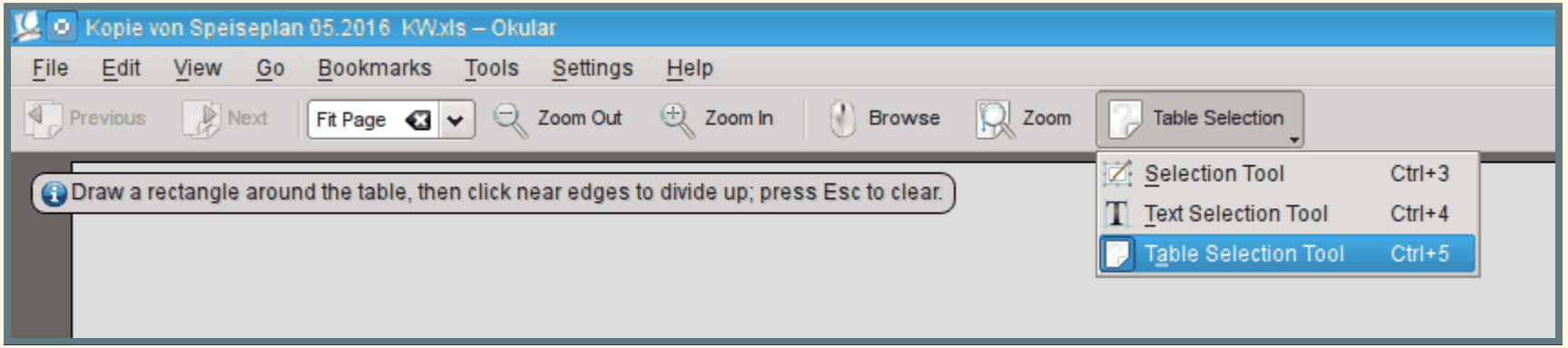
OCR

- PDF is vector based
- Endless zooming
- PNG with huge, well-defined letters
- Cut up into single cells
- We should OCR those easily (I hope)
- A path not taken

APOCALYPSE NOW – INTERACTIVE

OKULAR

- KDE document viewer
- Table selection tool


Kopie von Speiseplan 05.2016 KW.xls – Okular
File
Edit
View
Go
Bookmarks
Tools
Settings
Help
Previous
Next
Fit Page
Zoom Out
Zoom In
Browse
Zoom
Table Selection
Draw a rectangle around the table, then click near edges to divide up; press Esc to clear.
Selection Tool Ctrl+3
Text Selection Tool Ctrl+4
Table Selection Tool Ctrl+5

OKULAR

- KDE document viewer
- Table selection tool
- Recognizes row boundaries

Montag	Dienstag	Mittwoch	Donnerstag	Freitag
1. Februar 2016	2. Februar 2016	3. Februar 2016	4. Februar 2016	5. Februar 2016
Tomatencremesuppe	Gemüsebrühe[28]	Kartoffelsuppe[28]	Zwiebelsuppe	Erbsencremsuppe[26,28]
Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis	Riesencurrywurst[2,3,8] mit zweierlei Saucen[1,2] und Pommes Frites	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis[27] und Grillgemüse[28]	Gebackener Leberkäse[20w] mit Bohnengemüse[26] und Kartoffelstampf	Schlemmerfilet á la Bordelaise[4,20] mit Dillschaum[20,26] Tomatensalat und Kartoffeln
3,60 €	3,60 €	3,60 €	3,60 €	3,60 €
			Highlight der Woche	
Putengulasch mit Zucchini und Champignon[20w,26,28] dazu Nudeln[20w,23]	Kutterfrikadelle[4,21] Gurkensalat[2,3] und Kartoffelpüree	Hausgemachtes Schnitzel[20w] "Mailänder Art" mit Spaghetti[20w] und Tomatensauce	Hirschrückensteak an Rotweinjus, Rosenkohl und Butterspätzle[20w,]	Tanduri Masala Putenbraten auf Gemüse-Cous-Cous mit Joghurtdip[26]
3,60 €	3,60 €	3,60 €	5,90 €	3,60 €
Fit Menü	Fit Menü	Fit Menü	Fit Menü	Fit Menü
Gratinierte Maultaschen [20(W),23,28] auf Tomatenspinat	Frühlingsrolle[2,4,20,25,30] mit Süß-Saurem Dip und Basmatireis	Gemüselasagne[20w,26,28] mit Basilikum - Tomatensauce	Maisplinsen[20w,23,26] an gegrilltem Gemüse mit Limettensauce[26]	Rahmige Kartoffelsuppe[26,28] mit bunter Gemüseeinlage und Baguette[20,23,25,26,28,29,30,32]
3,30 €	3,30 €	3,30 €	3,30 €	3,30 €

Bei der Zubereitung der Speisen verwenden wir Allergie / Lebensmittelunverträglichkeit auslösende Stoffe als Zutaten, diese sind im Speiseplan gekennzeichnet.

OKULAR

- KDE document viewer
- Table selection tool
- Recognizes row boundaries
- Needs manual indication of column boundaries
- Does the heavy lifting
- Copy/paste as table data

Tomatencremesuppe	28 Gemüsebrühe	Kartoffelsuppe 28	Zwiebelsuppe	Erbsencremsuppe 26,28
Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis	2,3,8 Riesencurrywurst 1,2 mit zweierlei Saucen und Pommes Frites	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce 27 Nussreis 28 und Grillgemüse	20w Gebackener Leberkäse 26 mit Bohnengemüse und Kartoffelstampf	Schlemmerfilet á la 4,20 Bordelaise 20,26 mit DillschaumTomatensalat und Kartoffeln
3,60 €	3,60 €	3,60 €	3,60 €	3,60 €
			Highlight der Woche	
Putengulasch mit Zucchini 20w,26,28und Champignon dazu 20w,23 Nudeln	4,21 Kutterfrikadelle 2,3 Gurkensalat und Kartoffelpüree	20w Hausgemachtes Schnitzel "Mailänder Art" mit 20w Spaghetti und Tomatensauce	Hirschrückensteak an Rotweinjus, Rosenkohl und 20w, Butterspätzle	Tanduri Masala Putenbraten auf Gemüse-Cous-Cous mit 26 Joghurtdip
3,60 €	3,60 €	3,60 €	5,90 €	3,60 €
Fit Menü	Fit Menü	Fit Menü	Fit Menü	Fit Menü
Gratinierte Maultaschen 20(W),23,28 auf Tomatenspinat	2,4,20,25,30 Frühlingsrolle mit Süß-Saurem Dip und Basmatireis	20w,26,28 Gemüselasagne mit Basilikum - Tomatensauce	20w,23,26Maisplinsen an gegrilltem Gemüse mit Limettensauce26	26,28 Rahmige Kartoffelsuppemit bunter Gemüseeinlage und Baguette20,23,25,26,28,29,30,32
3,30 €	3,30 €	3,30 €	3,30 €	3,30 €

	A	B	C	D	E
1	Tomatencremesuppe	28 Gemüsebrühe	Kartoffelsuppe 28	Zwiebelsuppe	Erbsencremsuppe 26,28
2	Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis	2,3,8 Riesencurrywurst 1,2 mit zweierlei Saucen und Pommes Frites	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce 27 Nussreis 28 und Grillgemüse	20w Gebackener Leberkäse 26 mit Bohnengemüse und Kartoffelstampf	Schlemmerfilet á la 4,20 Bordelaise 20,26 mit DillschaumTomatensalat und Kartoffeln
3	3,60 €	3,60 €	3,60 €	3,60 €	3,60 €
4				Highlight der Woche	
5	Putengulasch mit Zucchini 20w,26,28und Champignon dazu 20w,23 Nudeln	4,21 Kutterfrikadelle 2,3 Gurkensalat und Kartoffelpüree	20w Hausgemachtes Schnitzel "Mailänder Art" mit 20w Spaghetti und Tomatensauce	Hirschrückensteak an Rotweinjus, Rosenkohl und 20w, Butterspätzle	Tanduri Masala Putenbraten auf Gemüse-Cous-Cous mit 26 Joghurtdip
6	3,60 €	3,60 €	3,60 €	5,90 €	3,60 €
7	Fit Menü	Fit Menü	Fit Menü	Fit Menü	Fit Menü
8	Gratinierte Maultaschen 20(W),23,28 auf Tomatenspinat	2,4,20,25,30 Frühlingsrolle mit Süß-Saurem Dip und Basmatireis	20w,26,28 Gemüselasagne mit Basilikum - Tomatensauce	20w,23,26Maisplinsen an gegrilltem Gemüse mit Limettensauce26	26,28 Rahmige Kartoffelsuppemit bunter Gemüseeinlage und Baguette20,23,25,26,28,29,30,32
9	3,30 €	3,30 €	3,30 €	3,30 €	3,30 €
10					

OTHER, SIMILAR SOLUTIONS

- `http://tabula.technology/`
  - *Tabula is a tool for liberating data tables locked inside PDF files.*
  - Used by i.e. `https://blog.openelections.net/`
- Online converters
  http://pdftables.com/ — https://online2pdf.com/pdf2excel — https://www.pdftoexcelonline.com/ — http://www.pdf2txt.de/ — etc.

PLATOON – ALL GUTS, NO GLORY

LOOK INSIDE PDF

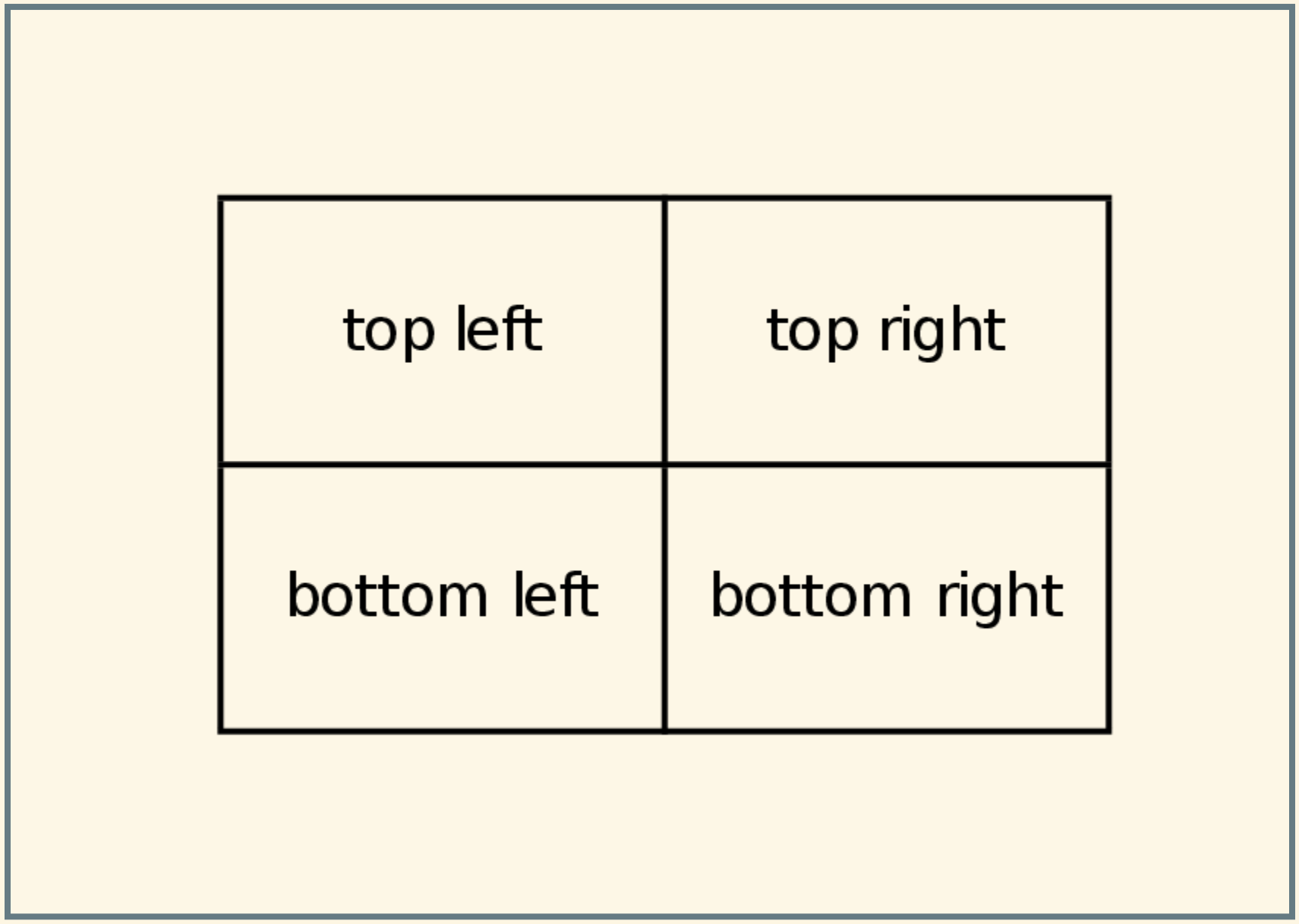
- PDF standard is ~~old~~ mature
- Versions from 1.0 to 1.7
- Basically human-readable
- ... but rarely in the wild
- Page dimensions defined by "MediaBox"
- Unit used is "pt" (point), 1/72 inch, 0.353 mm
- Coordinate system with 0/0 at lower left corner

MAKE IT (SLIGHTLY) READABLE

- `pdftk ... uncompress`
- Not really helpful with my PDF file
- QPDF does a much better job
- Still no real insight into problem domain

ELEMENTARY EXPERIMENTS

- Simple 2x2 table
- With cell borders
- Save as PDF
- Look at PDF source
- Revelation: cells don't really matter


top left
top right
bottom left
bottom right

[(t)18(o)12(p)36( )19(l)53(e)16(f)52(t)-3763(t)17(o)13(p)36( )18(r)37(i)53(g)36(h)35(t)]TJ

CPAN PDF MODULES

- Way too many!
- Some are old
- Some are unmaintained
- Some are one-off forks
- Look for alternatives

FULL METAL JACKET — THE TET OFFENSIVE

PDFLIB TET

- PDFlib: PDF developer toolbox
- Decades of Postscript and PDF experience
- Bindings for Perl, lots of other languagages
- Freely available evaluation version
- As well as extensive documentation
- TET = Text and Image Extraction Toolkit

TETML (YES, IT IS XML)

- Varying detail levels
- From word to single letters
- With bounding box etc.
- Reasonably good at detecting table structure

```
<Cell llx="109.16" lly="451.60" urx="467.89" ury="476.08">
    <Para>
        <Box llx="111.16" lly="462.28" ulx="111.16" uly="470.44" urx="604.93" ury
            <Line>
                <Text>Tomatencremesuppe Gemüsebrühe 28 Kartoffelsuppe 28</Text>
            </Line>
        </Box>
    </Para>
</Cell>
<Cell llx="554.68" lly="451.60" urx="604.93" ury="476.08">
    <Para>
        <Line>
            <Text>Zwiebelsuppe</Text>
        </Line>
    </Para>
</Cell>
<Cell llx="683.08" lly="451.60" urx="763.13" ury="476.08">
    <Para>
        <Box llx="683.08" lly="461.44" urx="763.13" ury="470.92">
            <Line>
                <Text>Erbsencremsuppe </Text>
            </Line>
            <Line>
                <Text>26,28</Text>
            </Line>
        </Box>
    </Para>
</Cell>
```

HOWEVER ...

- Evaluation version
- Should not be used in production
- Features and price (Linux: 1000 €) overkill
- Absolutely recommended, just not here

RTFM

CLOSER LOOK AT PDFTOTEXT (POPPLER)

- man page could have saved me days
- crop specific area of PDF
- `--layout` — maintain original physical layout
- WE HAVE A WINNER!
- **`pdftotext --layout`** does exactly what we want

Speiseplan KW 05

	Montag 1. Februar 2016	Dienstag 2. Februar 2016	Mittwoch 3. Februar 2016	Donnerst 4. Februa
Suppe	Tomatencremesuppe	Gemüsebrühe 28	Kartoffelsuppe 28	Zwiebelsupp
	0,85 EUR			
Haupt-gerichte	Koreanisches Feuerfleisch vom Schwein mit Mangold und Wildreis 3,60 EUR	Riesencurrywurst2,3,8 mit zweierlei Saucen1,2 und Pommes Frites 3,60 EUR	Gebratene Hähnchenbrust in Rosmarin-Tomatensauce Nussreis27 und Grillgemüse28 3,60 EUR	Gebackener Leberk mit Bohnengemüse2 Kartoffelsta 3
	Putengulasch mit Zucchini und Champignon20w,26,28 dazu Nudeln20w,23 3,60 EUR Fit Menü	Kutterfrikadelle4,21 Gurkensalat2,3 und Kartoffelpüree 3,60 EUR Fit Menü	Hausgemachtes Schnitzel20w "Mailänder Art" mit Spaghetti20w und Tomatensauce 3,60 EUR Fit Menü	Highlight der Hirschrückenste Rotweinjus, Rosen Butterspätzl 5 Fit Menü
Vegeta-risch	Gratinierte Maultaschen 20(W),23,28 auf Tomatenspinat 3,30 EUR	Frühlingsrolle2,4,20,25,30 mit Süß-Saurem Dip und Basmatireis 3,30 EUR	Gemüselasagne20w,26,28 mit Basilikum - Tomatensauce 3,30 EUR	Maisplinsen20w gegrilltem Gemü Limettensauc 3
Salate	Bei der Zubereitung der Speisen verwenden wir Allergie / Lebensmittelunverträglichkeit auslösende Stoffe als Zutaten, Trotz größtmöglicher Sorgfalt können wir jedoch nicht ausschließen, dass technisch unvermeidbare Einträge von Allergie / Leben ggf. weitere mögliche Einträge auch in andere Gerichte gelangen. Weitere Auskünfte dazu erteilt Ihnen g			

	Wildreis	Pommes Frites	Nussreis	Kartoffelst
Beilagen	Nudeln	Kartoffelpüree	Spaghetti	Spätzle
0,70 EUR		Basmatireis		
	Tagesgemüse28	Tagesgemüse28	Tagesgemüse28	Tagesgemüse2

Dessert
Groß 0,80 EUR
Klein 0,50 EUR
Zusatzstoffe sind mit Zahlen gekennzeichnet siehe Aushang
Bei Fragen zu Dussmann können Sie auf www.dussmann.com klicken!

Speiseplanänderungen vorbehalten

THANK YOU!

- Slides will be linked on conference page
- Questions?