ORPALIS
A PSPDFKit Company

How document understanding can leverage your PDF workflow

How to overcome the different challenges - fields of application

working together!

PDF Days Europe 2022 | Berlin

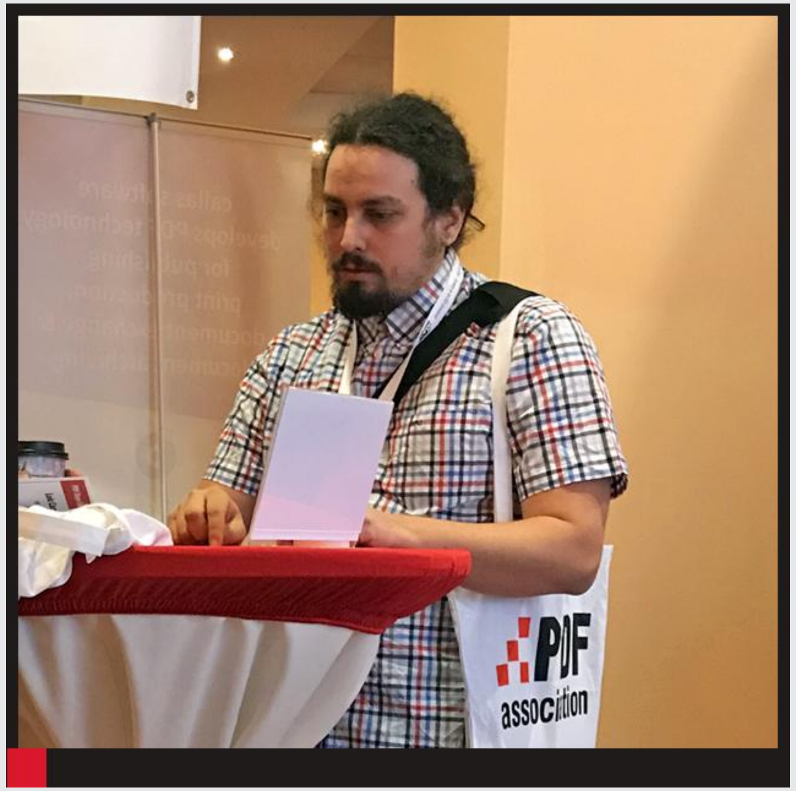
Content

- PDF, an Unstructured Format
- Challenges of PDF
- What is Document Understanding?
- Document Layout Analysis
- Optical Character Recognition
- Key-Value Pair
- Natural Language Processing
- Named-Entity Recognition
- Fields of Application

Speakers

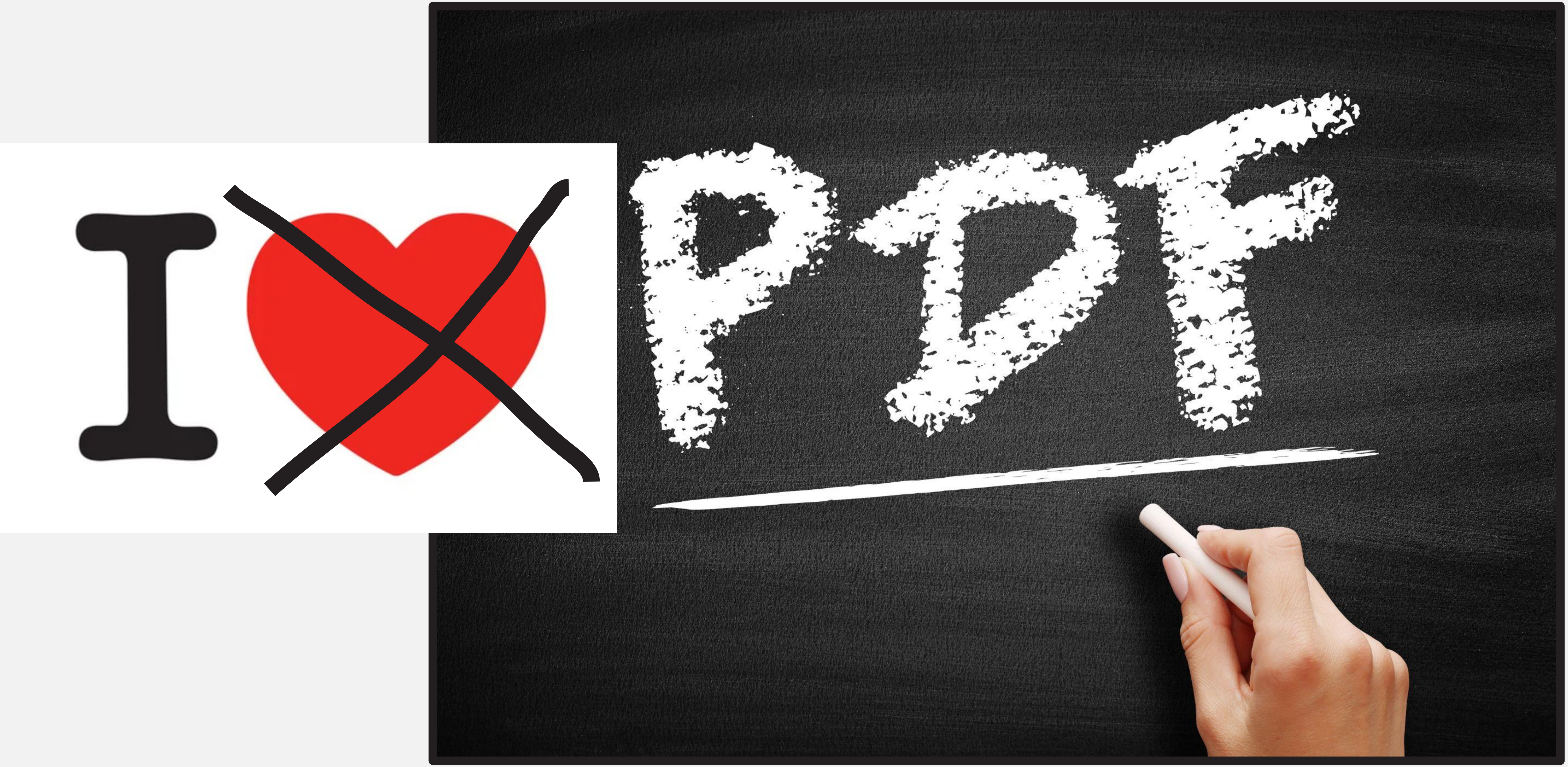
Matúš Pizúr

Senior developer and PDF specialist



Elodie Tellier

Copywriter and PDF Association Board Member



I ♥
PDF

PDF Introduction

File format for digital distribution of final form documents

Created with focus on:

- Consistency
- Fidelity
- Convenience
- Security

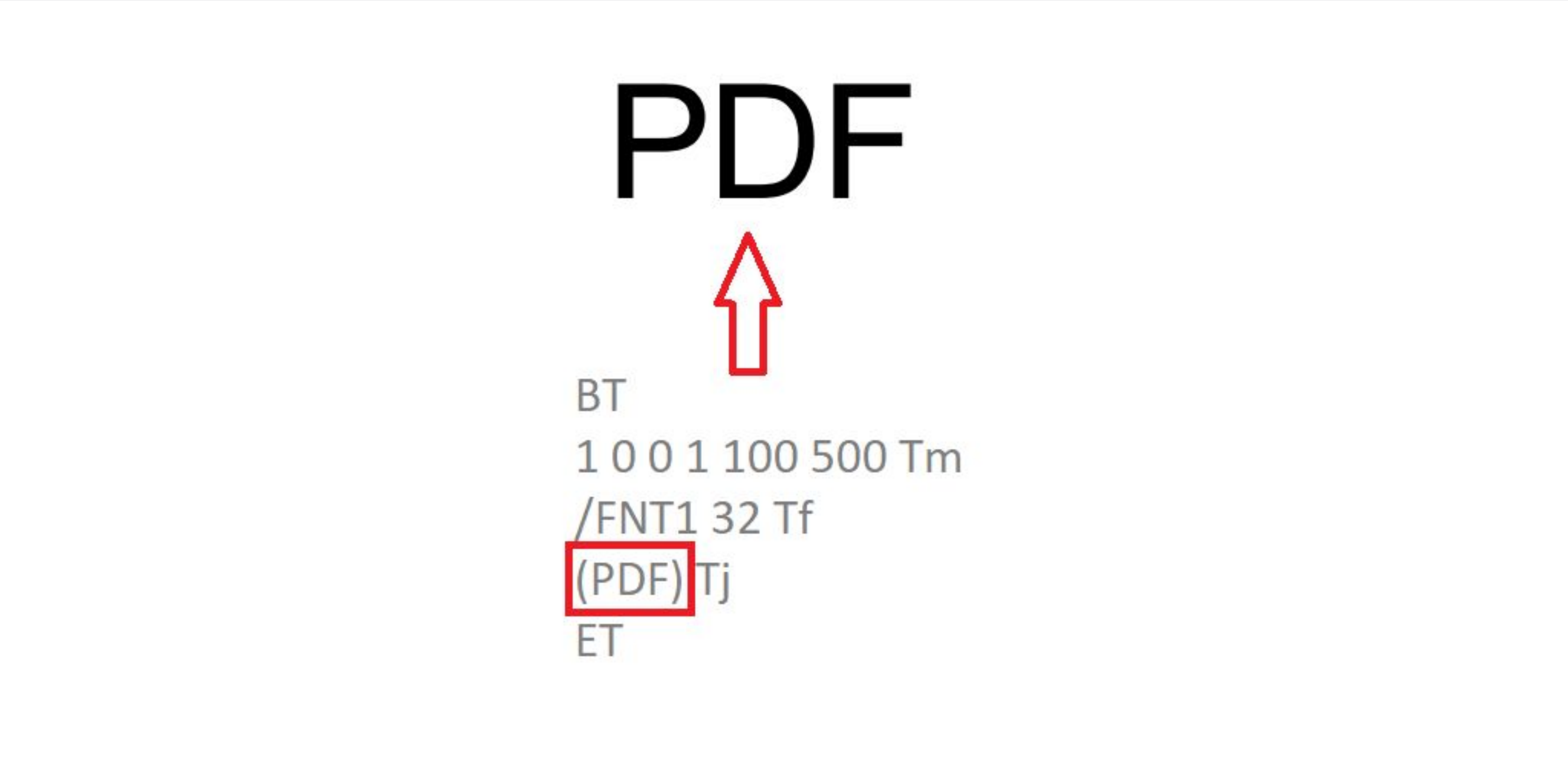
PDF Composition: Digitally-Born PDF

Product of generation from data or conversion of other digital documents

- Contain unstructured text and images as individual graphic objects
- Makes use of all available features of PDF specification
- No structure on the content stream level

PDF Composition: Digitally-Born PDF

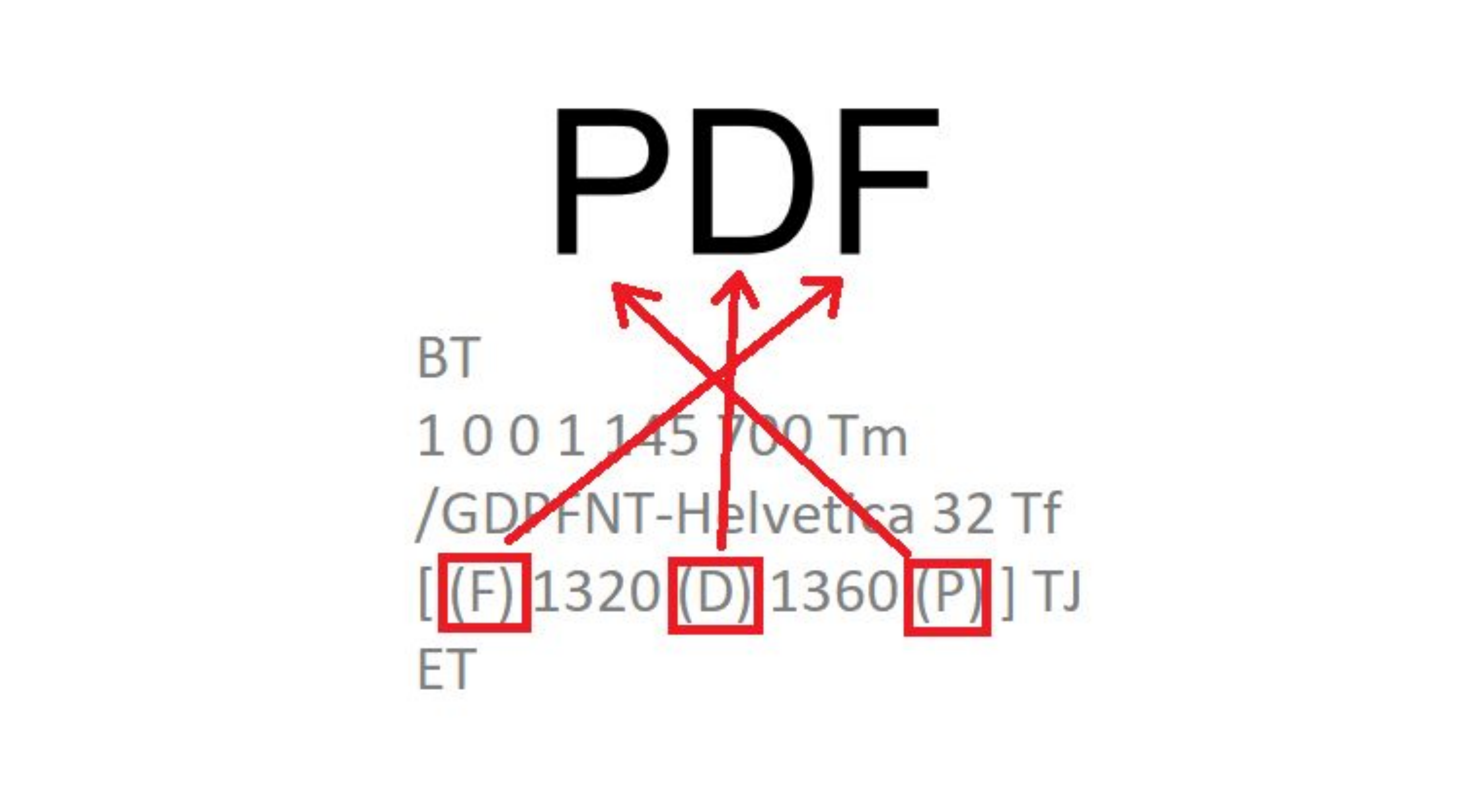
Basic example



```
BT
1 0 0 1 100 500 Tm
/FNT1 32 Tf
(PDF) Tj
ET
```

PDF Composition: Digitally-Born PDF

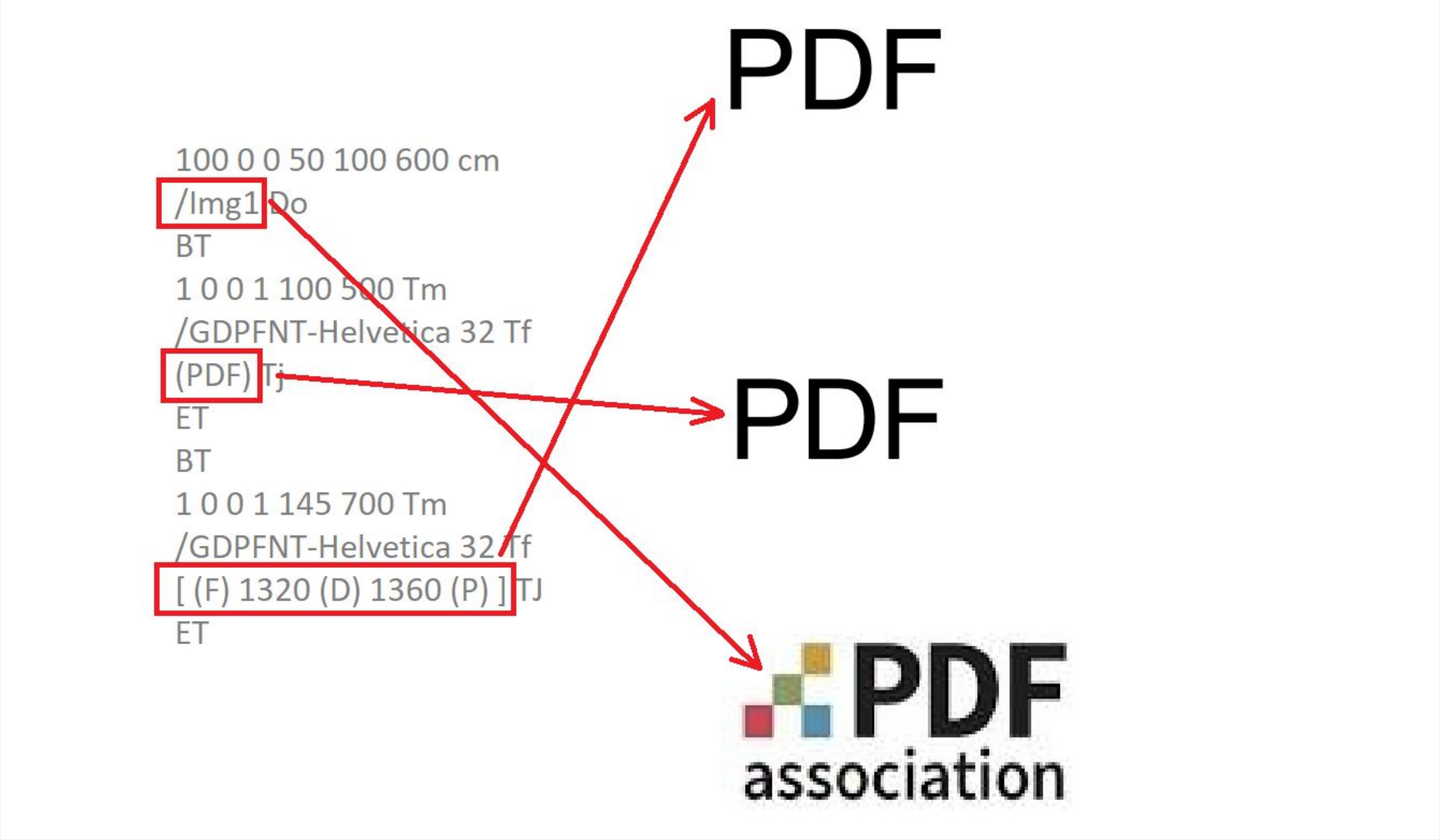
Not so basic



```
BT
1 0 0 1 145 700 Tm
/GDPFNT-Helvetica 32 Tf
[(F) 1320 (D) 1360 (P)] TJ
ET
```

PDF Composition: Digitally-Born PDF

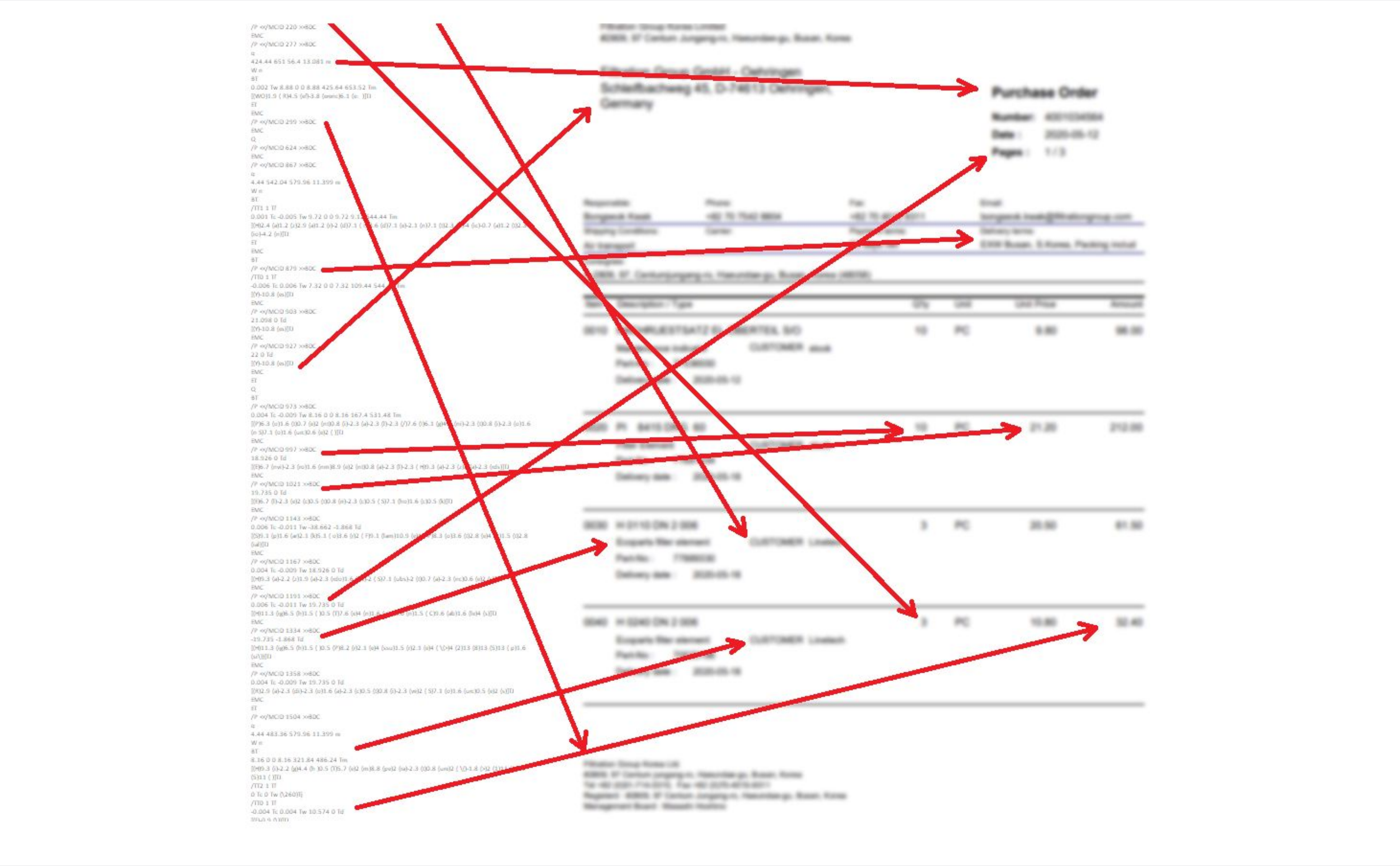
Object level



```
100 0 0 50 100 600 cm
/Img1 Do
BT
1 0 0 1 100 500 Tm
/GDPFNT-Helvetica 32 Tf
(PDF) Tj
ET
BT
1 0 0 1 145 700 Tm
/GDPFNT-Helvetica 32 Tf
[ (F) 1320 (D) 1360 (P) ] TJ
ET
```

PDF Composition: Digitally-Born PDF

…



Purchase Order

Number:

Date : 2020-05-12

Pages : 1 / 3

Challenges of PDF: Digitally-Born PDF

Tagging and logical structure

- Tagged PDF introduced in PDF 1.4 (2001)
- PDF/UA - ISO 14289 (2012)
- Still only small portion of consumed files are tagged (14% based on 2018 data)

Challenges of PDF: Digitized PDF

- Single image per page
- No structure
- Constitutes a considerable portion of all PDF documents consumed

```
395.75999 0 0 611.75999 0 0 cm
/Im0 Do
```

INVOICE

WE APPRECIATE YOUR BUSINESS!!

PDF Without Structure

working together!

Unstructured documents represent 90% of all documents generated

These documents can be:

- Not searchable and discoverable
- Problematic in day to day work
- Unable to be processed in automated workflows
- Not accessible

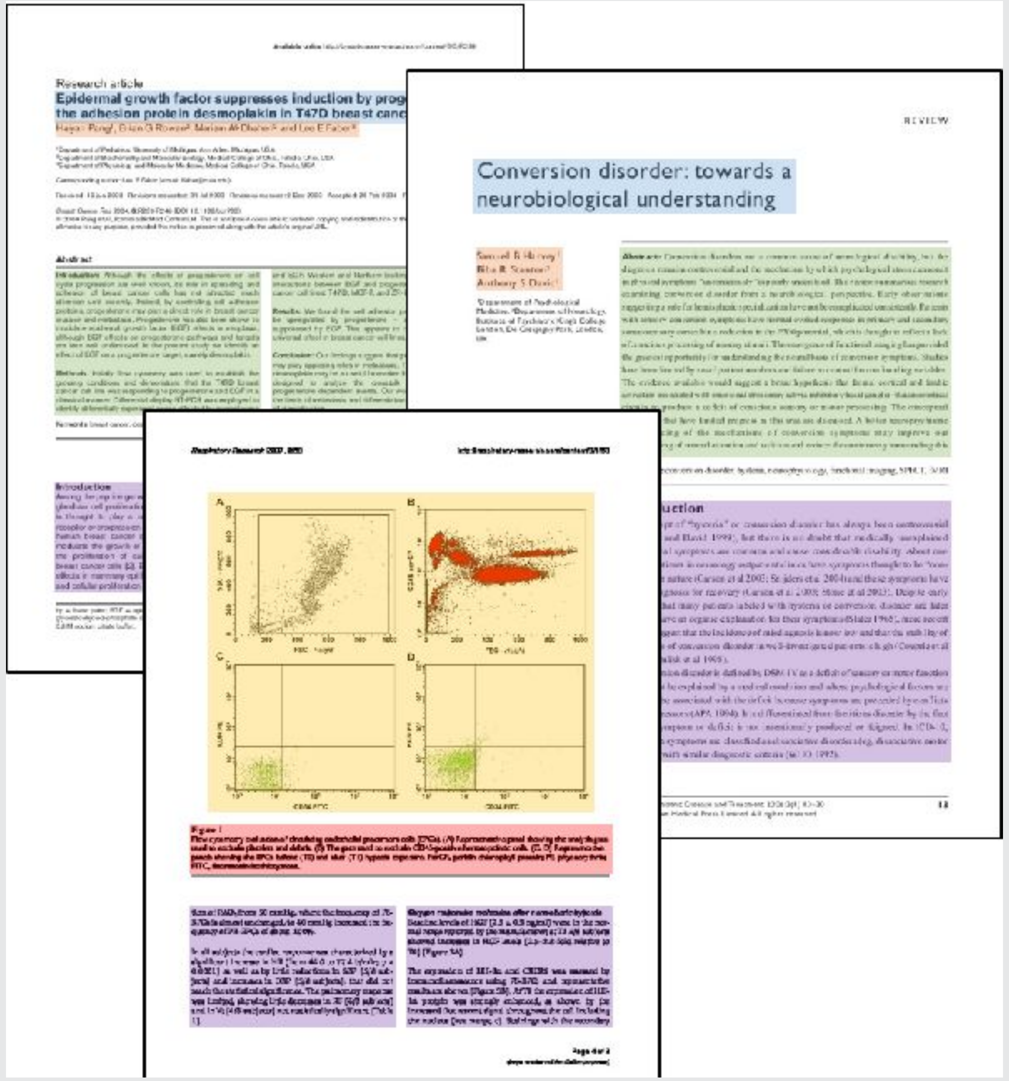
What is Document Understanding?

Document understanding is at the core of any Intelligent Document Processing solution

- Layout analysis + recognition → First level of document understanding
- First level of document understanding + NLP → Reinforced document understanding

Tech 1: Document Layout Analysis (DLA)

- DLA is the identification and categorization of regions
- DLA implies a geometric analysis of tables, pictures, equations, and barcodes and a logical layout analysis (paragraphs, lines, words, characters) of the document



Research article

Epidermal growth factor suppresses induction by prog

the adhesion protein desmoplakin in T47D breast canc

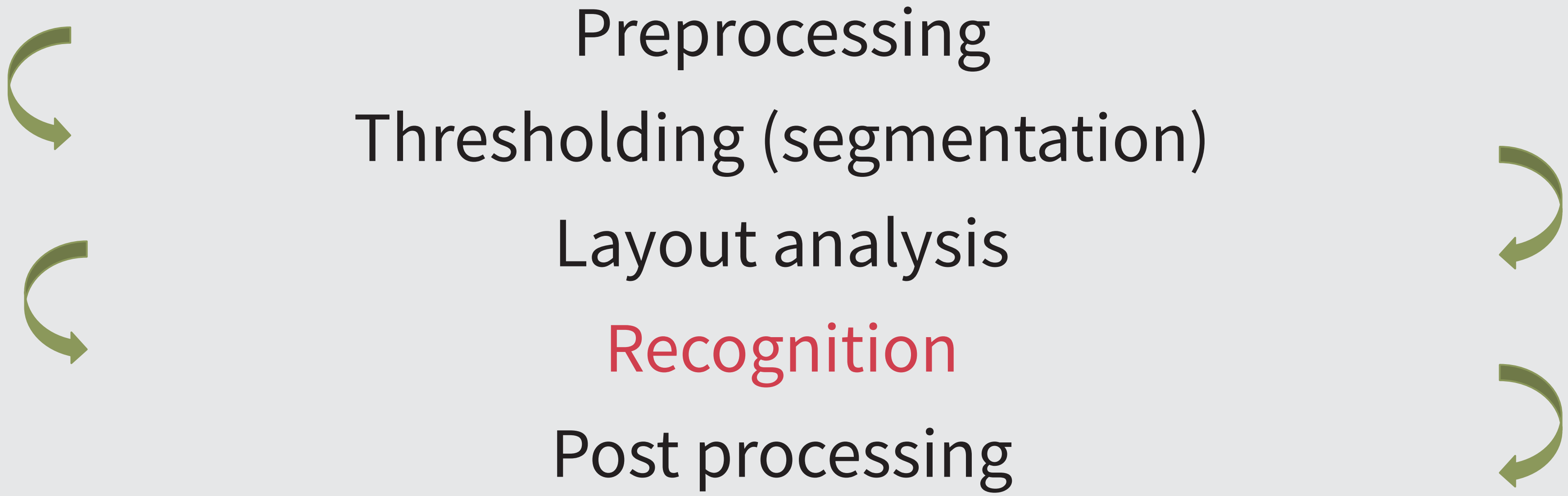
REVIEW

Conversion disorder: towards a neurobiological understanding

Soto, C. and Shinjae Yoo. "Visual Detection with Context for Document Layout Analysis." EMNLP (2019).

Tech 2: Optical Character Recognition (OCR)

What’s in an OCR engine?



Preprocessing

Thresholding (segmentation)

Layout analysis

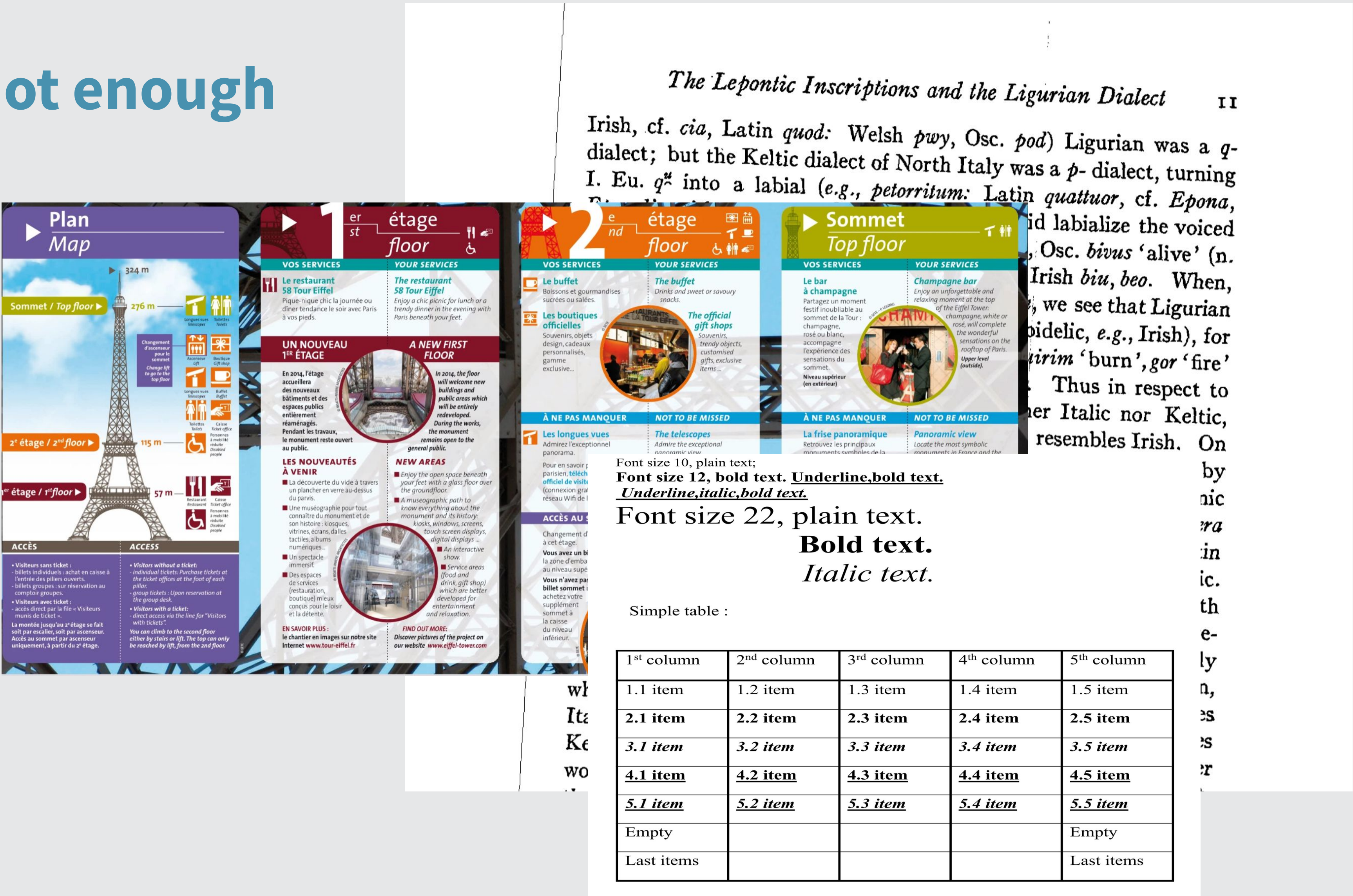
Recognition

Post processing

The Issues with Traditional OCR

A traditional/standard OCR is not enough

- colored backgrounds
- glaring
- skew
- text in tables and graphics
- handwritten text
- same font in different size
- underlined text
- hard to scale



324 m
Sommet / Top floor ▶
276 m
Longues vues Telescopes
Toilettes Toilets
Changement d'ascenseur pour le sommet
Change lift to go to the top floor
Ascenseur Lift
Boutique Gift shop
Longues vues Telescopes
Buffet Buffet
Toilettes Toilets
Caisse Ticket office
2e étage / 2nd floor ▶
115 m
Personnes à mobilité réduite Disabled people
1er étage / 1st floor ▶
57 m
Restaurant Restaurant
Caisse Ticket office
Personnes à mobilité réduite Disabled people
© SETE
© MOATTI-RIVIERE ARCHITECTES
RESTAURANTS DE LA TOUR EIFFEL
© SETE – P. LESTANG
CHAMP

Plan
Map

ACCÈS

- **Visiteurs sans ticket :**
  - billets individuels : achat en caisse à l'entrée des piliers ouverts.
  - billets groupes : sur réservation au comptoir groupes.
- **Visiteurs avec ticket :**
  - accès direct par la file « Visiteurs munis de ticket ».

La montée jusqu'au 2e étage se fait soit par escalier, soit par ascenseur. Accès au sommet par ascenseur uniquement, à partir du 2e étage.

ACCESS

- ***Visitors without a ticket:***
  - *individual tickets: Purchase tickets at the ticket offices at the foot of each pillar.*
  - *group tickets : Upon reservation at the group desk.*
- ***Visitors with a ticket:***
  - *direct access via the line for "Visitors with tickets".*

You can climb to the second floor either by stairs or lift. The top can only be reached by lift, from the 2nd floor.

1er étage
1st floor

VOS SERVICES

Le restaurant 58 Tour Eiffel
Pique-nique chic la journée ou dîner tendance le soir avec Paris à vos pieds.

YOUR SERVICES

The restaurant 58 Tour Eiffel
Enjoy a chic picnic for lunch or a trendy dinner in the evening with Paris beneath your feet.

UN NOUVEAU 1ER ÉTAGE

En 2014, l'étage accueillera des nouveaux bâtiments et des espaces publics entièrement réaménagés. Pendant les travaux, le monument reste ouvert au public.

LES NOUVEAUTÉS À VENIR

- La découverte du vide à travers un plancher en verre au-dessus du parvis.
- Une muséographie pour tout connaître du monument et de son histoire : kiosques, vitrines, écrans, dalles tactiles, albums numériques...
- Un spectacle immersif.
- Des espaces de services (restauration, boutique) mieux conçus pour le loisir et la détente.

EN SAVOIR PLUS :
le chantier en images sur notre site Internet www.tour-eiffel.fr

A NEW FIRST FLOOR

In 2014, the floor will welcome new buildings and public areas which will be entirely redeveloped. During the works, the monument remains open to the general public.

NEW AREAS

- *Enjoy the open space beneath your feet with a glass floor over the groundfloor.*
- *A museographic path to know everything about the monument and its history: kiosks, windows, screens, touch screen displays, digital displays ...*
- *An interactive show.*
- *Service areas (food and drink, gift shop) which are better developed for entertainment and relaxation.*

FIND OUT MORE:
Discover pictures of the project on our website www.eiffel-tower.com

2e étage
2nd floor

VOS SERVICES

Le buffet
Boissons et gourmandises sucrées ou salées.

Les boutiques officielles
Souvenirs, objets design, cadeaux personnalisés, gamme exclusive...

YOUR SERVICES

The buffet
Drinks and sweet or savoury snacks.

The official gift shops
Souvenirs, trendy objects, customised gifts, exclusive items ...

À NE PAS MANQUER

Les longues vues
Admirez l'exceptionnel panorama.

Pour en savoir p
parisien, téléch
officiel de visite
(connexion grat
réseau Wifi de l

ACCÈS AU S

Changement d'
à cet étage.

Vous avez un bi
la zone d'emba
au niveau supé

Vous n'avez pas billet sommet :
achetez votre supplément sommet à la caisse du niveau inférieur.

NOT TO BE MISSED

The telescopes
Admire the exceptional panoramic view

Sommet
Top floor

VOS SERVICES

Le bar à champagne
Partagez un moment festif inoubliable au sommet de la Tour : champagne, rosé ou blanc, accompagne l'expérience des sensations du sommet.
Niveau supérieur (en extérieur)

YOUR SERVICES

Champagne bar
Enjoy an unforgettable and relaxing moment at the top of the Eiffel Tower: champagne, white or rosé, will complete the wonderful sensations on the rooftop of Paris.
Upper level (outside).

À NE PAS MANQUER

La frise panoramique
Retrouvez les principaux monuments symboles de la

NOT TO BE MISSED

Panoramic view
Locate the most symbolic monuments in France and the

The Lepontic Inscriptions and the Ligurian Dialect 11

Irish, cf. *cia*, Latin *quod:* Welsh *pwy*, Osc. *pod*) Ligurian was a *q*-
dialect; but the Keltic dialect of North Italy was a *p*- dialect, turning
I. Eu. q^{u} into a labial (*e.g.*, *petorritum:* Latin *quattuor*, cf. *Epona*,
id labialize the voiced
, Osc. *bivus* 'alive' (n.
Irish *biu*, *beo*. When,
, we see that Ligurian
oidelic, *e.g.*, Irish), for
irim 'burn', *gor* 'fire'
Thus in respect to
er Italic nor Keltic,
resembles Irish. On
by
nic
ra
in
ic.
th
e-
ly
n,
es
es
er

wh
Ita
Ke
wo

Font size 10, plain text;
Font size 12, bold text. Underline,bold text.
Underline,italic,bold text.

Font size 22, plain text.

Bold text.

Italic text.

Simple table :

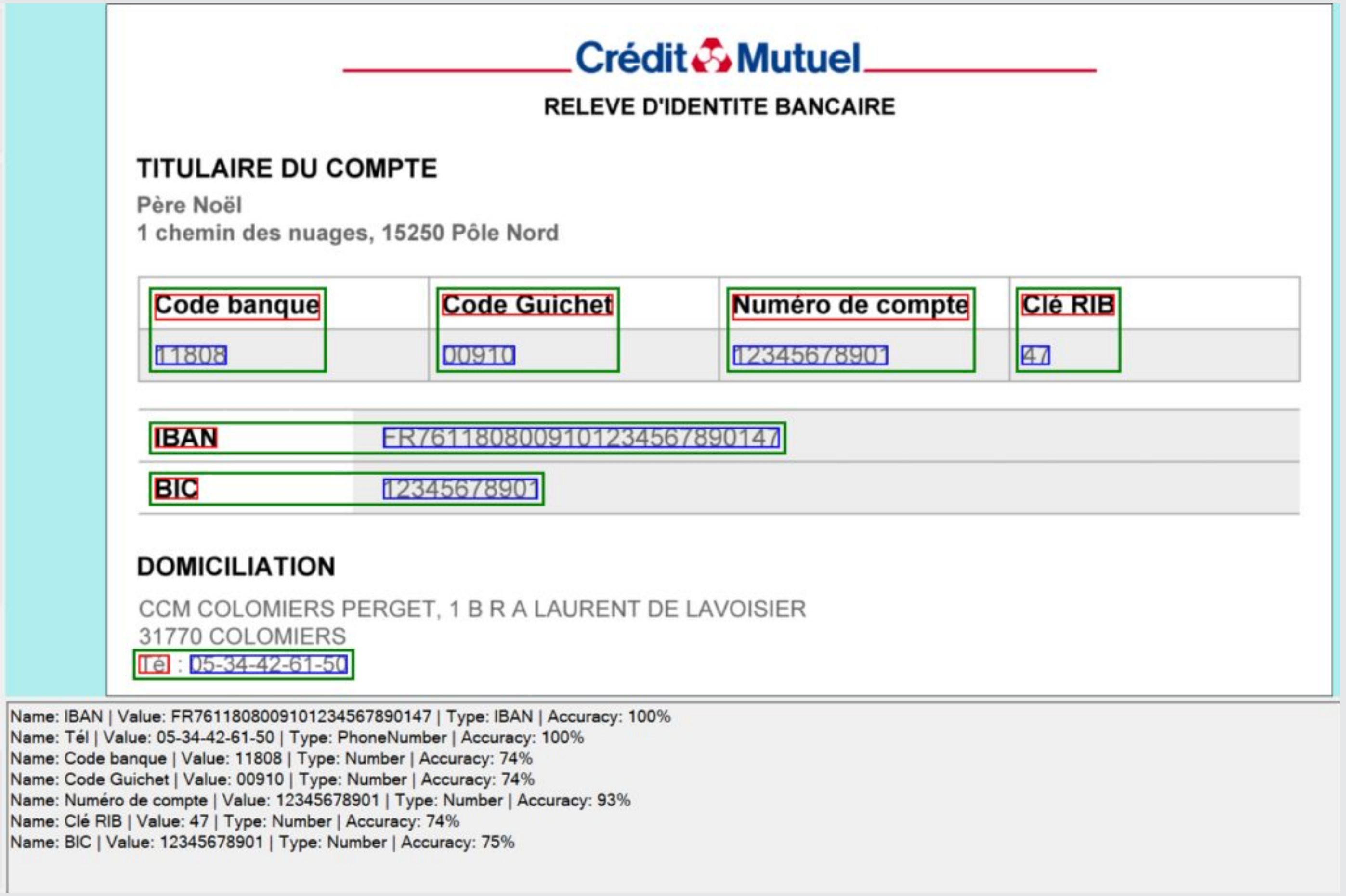
1st column	2nd column	3rd column	4th column	5th column
1.1 item	1.2 item	1.3 item	1.4 item	1.5 item
2.1 item	**2.2 item**	**2.3 item**	**2.4 item**	**2.5 item**
3.1 item	***3.2 item***	***3.3 item***	***3.4 item***	***3.5 item***
4.1 item	**4.2 item**	**4.3 item**	**4.4 item**	**4.5 item**
5.1 item	***5.2 item***	***5.3 item***	***5.4 item***	***5.5 item***
Empty				Empty
Last items				Last items

Tech 3: Textual Content Key-Value Association

Key-Value Pair (KVP)

KVPs are two related data items, a key, and a value. The key defines the data and is fixed, and the value is variable and describes the key.

Example: BIC (key) : 12345678901 (value)



Crédit Mutuel

RELEVE D'IDENTITE BANCAIRE

TITULAIRE DU COMPTE

Père Noël
1 chemin des nuages, 15250 Pôle Nord

Code banque	Code Guichet	Numéro de compte	Clé RIB
11808	00910	12345678901	47

IBAN	FR7611808009101234567890147
BIC	12345678901

DOMICILIATION

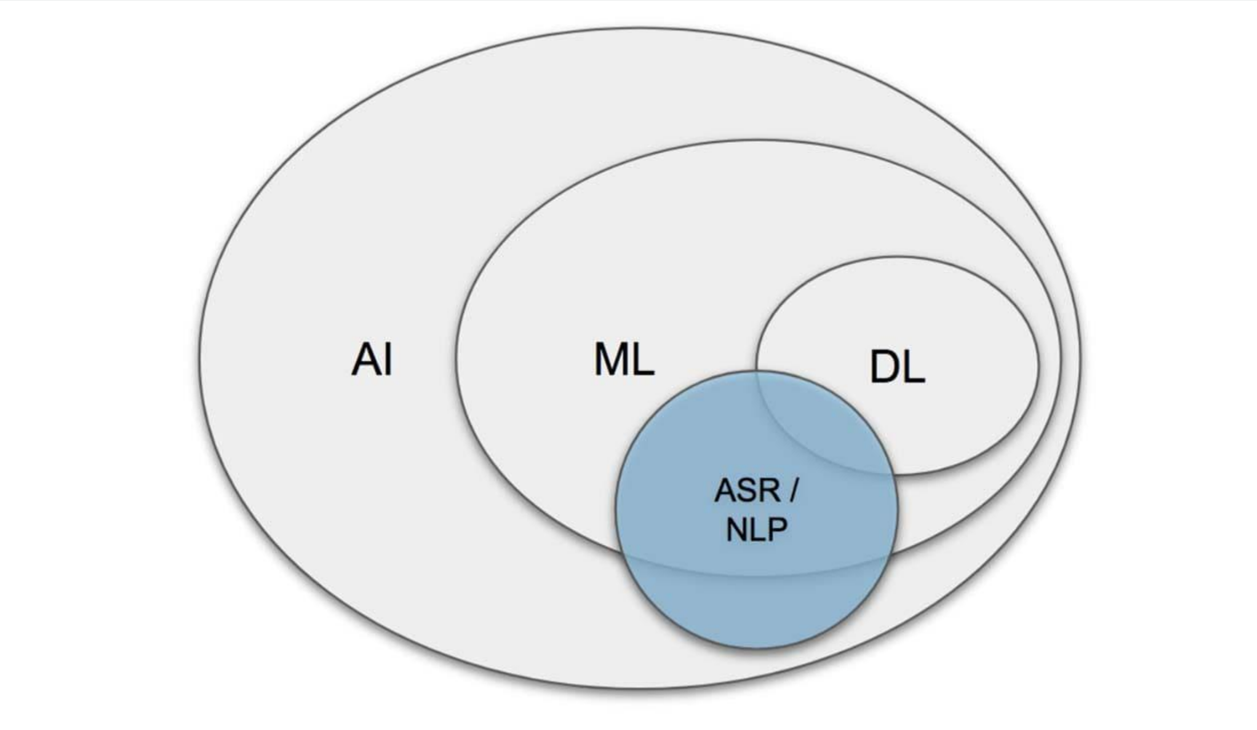
CCM COLOMIERS PERGET, 1 B R A LAURENT DE LAVOISIER
31770 COLOMIERS
Tél : 05-34-42-61-50

```
Name: IBAN | Value: FR7611808009101234567890147 | Type: IBAN | Accuracy: 100%
Name: Tél | Value: 05-34-42-61-50 | Type: PhoneNumber | Accuracy: 100%
Name: Code banque | Value: 11808 | Type: Number | Accuracy: 74%
Name: Code Guichet | Value: 00910 | Type: Number | Accuracy: 74%
Name: Numéro de compte | Value: 12345678901 | Type: Number | Accuracy: 93%
Name: Clé RIB | Value: 47 | Type: Number | Accuracy: 74%
Name: BIC | Value: 12345678901 | Type: Number | Accuracy: 75%
```

Tech 4: Natural Language Processing (NLP)

working together!

Data extraction



1. Extracting unstructured information or text with OCR or OCR + DL

2. Making sense of this unstructured information with DL + NLP

Tech 5: Named-Entity Recognition (NER)

- NER is a form of Natural Language Processing (NLP), a subfield of artificial intelligence
- NER helps with KVP extraction: faster & more accurate
- Building an algorithm with NER is not sufficient to address all the challenges

MAKE
THINGS
MUCH
BETTER

Fields of Application: Improved OCR

Benefits of a performant document understanding system on OCR

- improves the OCR recognition step, especially with LSTM-based recognizers
- permits to move beyond simple OCR processes and towards document understanding

Fields of Application: Tagging

Benefits of a performant document understanding system on tagging

- automatic tagging
- better accessibility support by making the conversion to PDF/UA easier

Fields of Application: Conversion

Benefits of a performant document understanding system on conversion

- structured layout conversion
- improves the conversion of a fixed into an editable layout (PDF to Office formats)

Fields of Application: Indexing & Labeling

Benefits of a performant document understanding system on automation processes

- automatic indexing
- automatic labeling

Fields of Application: Redaction

Benefits of a performant document understanding system on redaction

- automatic redaction

Fields of Application: Hyper-Compression

Benefits of a performant document understanding system on compression

- better MRC compression
- better color detection compression
- better PDF optimization

?

QUESTIONS

Keep in touch with us! m.pizur@orpalis.com / e.tellier@orpalis.com

www.orpalis.com