Extracting Info from Company Statutes

BeCode Use Case - September 2019

Introduction

The Team

Stany Boes

Director

Jeroen Bolle

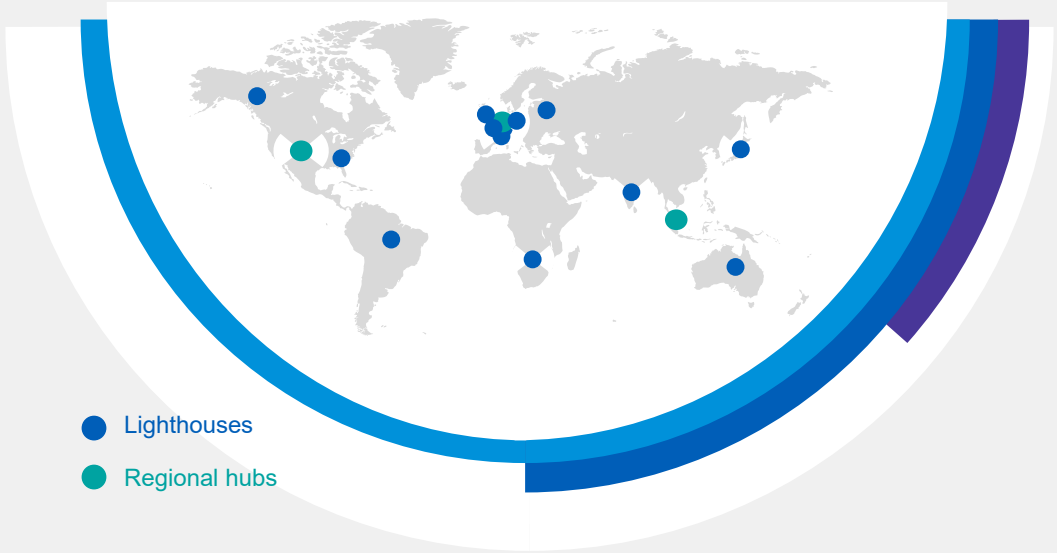
Advisor

Jonas Vanden Branden

Advisor

KPMG Lighthouse

Center of excellence for data-driven technologies



The KPMG Lighthouse network combines our data-driven technologies and capabilities with our deep-rooted domain expertise to accelerate innovation, drive speed and relevance and ensure global scale for data-driven solutions.
KPMG Lighthouse teams leverage data, analytics and artificial intelligence technologies to build and deliver solutions that transform the business of our clients.

3bn+
USD global revenue

12,500+
KPMG experts around the world

7,000+
delivered client engagements per year

1,700+
data scientists

600+
pre-built solutions

8
Insights Centers

4
strategic partnerships with Google, IBM, Microsoft and Oracle

3
global platforms

- KPMG Ignite
- KPMG Sofy
- KPMG Signals Repository

Advanced data management
Data engineering
Data mining
Big data

Data and analytics

Data visualization
Smart data transformation
Analytical modelling
Analytical enterprise
Advanced analytics
Deep learning
Algorithm assurance
Pattern recognition

Intelligent Automation

Virtual agents
Cognitive automation

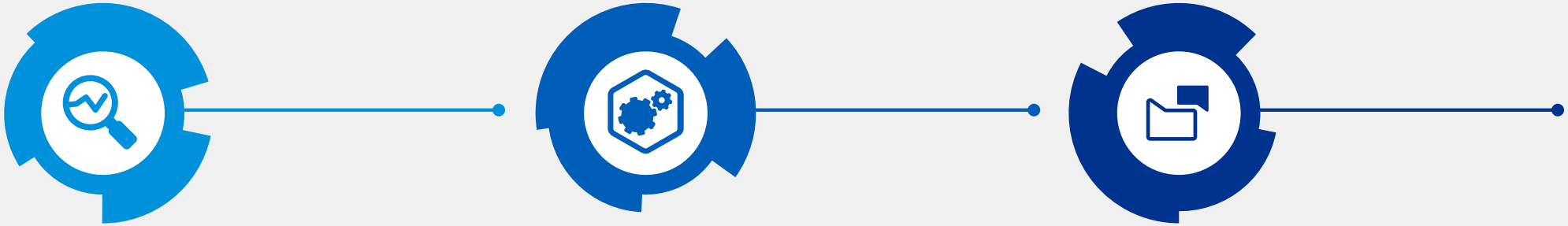
Artificial Intelligence

Optimization and simulation
Knowledge-based systems
Natural language processing
Voice / image recognition
Reasoning
Machine learning
Robotic process automation
Decision modelling

KPMG

KPMG Lighthouse teams

The Global Lighthouse provides the access point to 12,500 professionals across the globe



Data scientists

Strong experience in analytics, statistics, data mining, machine learning, natural language processing and/or mathematics.

Problem-solving ability through the use and/or development of algorithms, models, testing, etc.

Generally MS or PhD-level math, statistics, or engineering.

Software engineers

Strong experience with large scale and/or distributed processing methodologies such as Hadoop, Storm, Spark, and many others.

Sophisticated ability to rapidly ingest, transform and mine data.

Ability to evaluate, design, build, test and manage ‘big data’ architectures.

Consultants

Strong business consulting acumen and statistical background, combined with real-world experience in applying analytics to solve business issues.

Practical understanding of advanced analytics methods and ‘big data’ software; client solutioning expertise.

KPMG Ignite

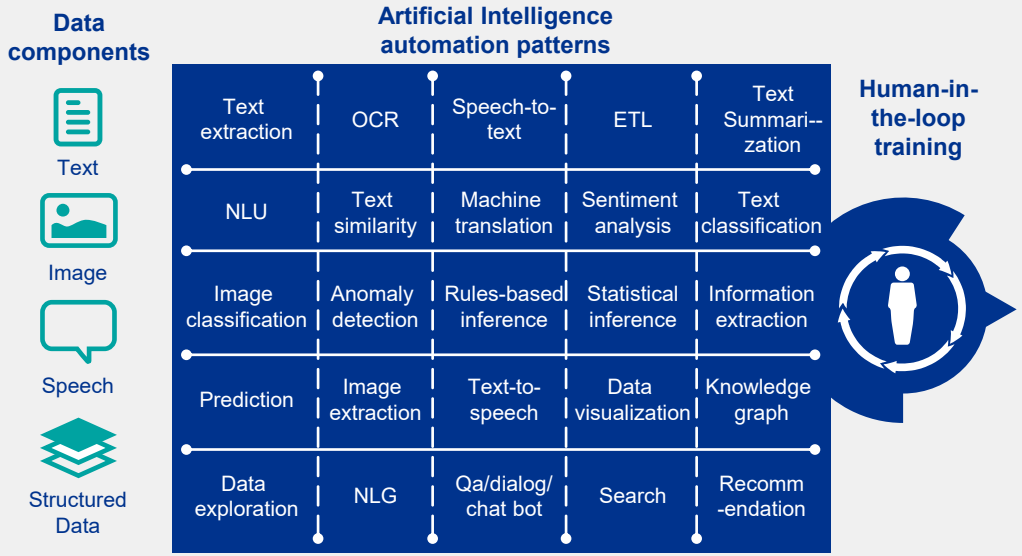
KPMG's portfolio of Artificial Intelligence capabilities

What it is…

KPMG Ignite is the KPMG's portfolio of artificial intelligence capabilities. It includes domain expertise, integrated open source tools and frameworks, strategic technology partnerships, KPMG-developed IP, frameworks and patterns, as well as research and experimentation

KPMG Ignite ingests various types of data components from different sources and applies AI-based automation patterns to create intelligent workflows to solve business problems. The patterns each cover an AI capability, from sentiment analysis, text classification and image to text.

Platform features



Exemplary solutions

Cognitive contract analytics
Use RPA and NLP to evaluate compliance

Chat/ voice bots
Automated handling of frequent customer requests

Call Center Analytics
Natural Language Processing-based solution for call volume forecasting and increased customer satisfaction

Value delivered

- Increases accuracy through 100 percent coverage versus traditional sampling approaches
- Reduces cost and development time needed to produce insights
- Enables humans to be precise and manage consistency
- Leverages the knowledge and experience of the very best subject matter experts
- Increases transparency through audit logs that show how data has been processed

KPMG Signals Repository

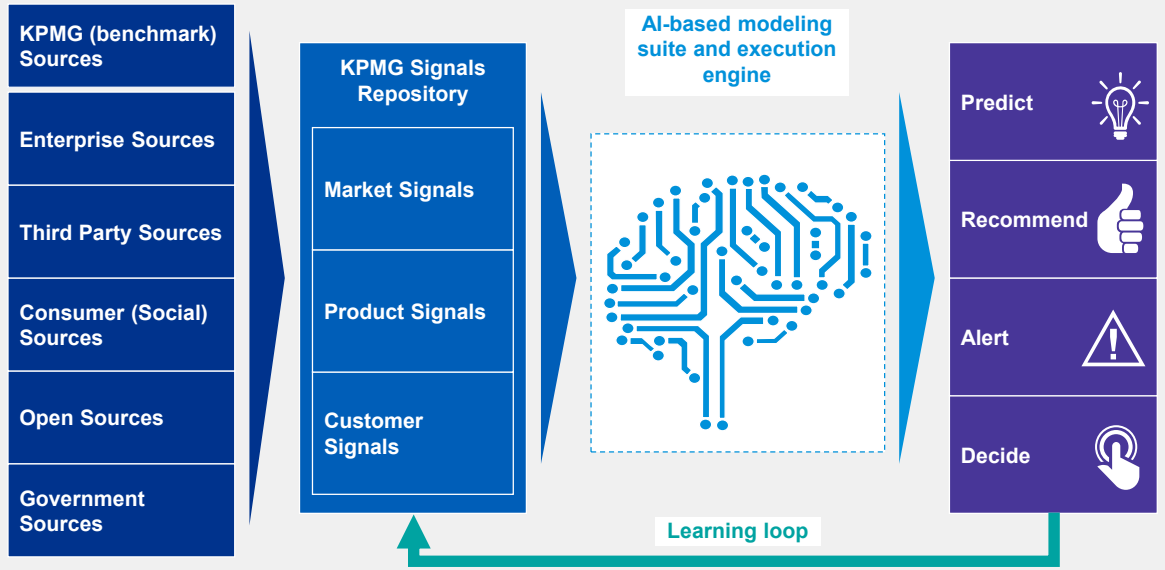
KPMG's big data and decision science platform

What it is…

The KPMG Signals Repository is an active platform that continuously harvests a broad variety of signals from Public and Private sources. Based on latest decision science, it effectively creates a Big Data Fabric from exogenous and endogenous data that can be used by AI and machine learning technologies to drive improved decisions and actions.

Within the Signals Repository, structured and unstructured data is transformed into complex expressions, then subsequently engineered into features within models. The breadth and specificity of Signals drive unprecedented accuracy in predictions and business execution outcomes.

Platform features



Exemplary solutions

Pricing
Determine price combinations that drive greatest participation and profit.

Point of Sales optimization
Use Big Data and Advanced Analytics to determine the unique drivers of demand

Customer satisfaction
Listen to the voice of customers from their digital footprint to understand what drives satisfaction

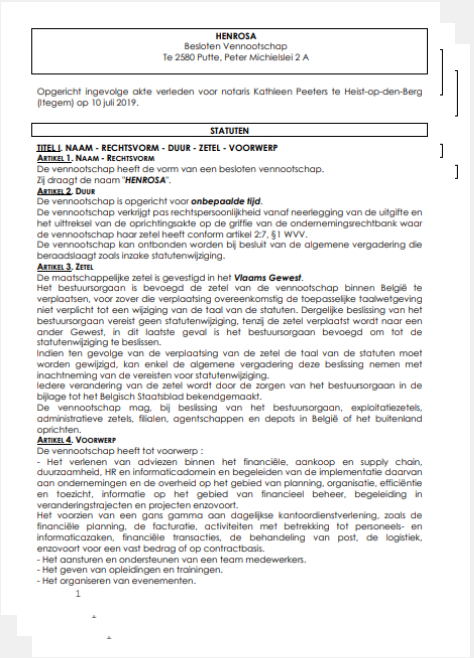
Employee retention
Understand key drivers of employee retention to predict employee churn rates

Value delivered

- More than 10,000 signals of traditional and non-traditional data, like "SoLoMo" (Social, Local, Mobile)
- Signals are continually monitored for changes in quality and impact
- Automated, self-serve access to the created compendium of signals to improve business decisions
- Leverages Big Data to augment or automate Decision-making and generate material business results

The Challenge

Goal



HENROSA
Besloten Vennootschap
Te 2580 Putte, Peter Michielslei 2 A

Opgericht ingevolge akte verleden voor notaris Kathleen Peeters te Heist-op-den-Berg (Hegem) op 10 juli 2019.

STATUTEN

TITEL I. NAAM - RECHTSVORM - DUUR - ZETEL - VOORWERP

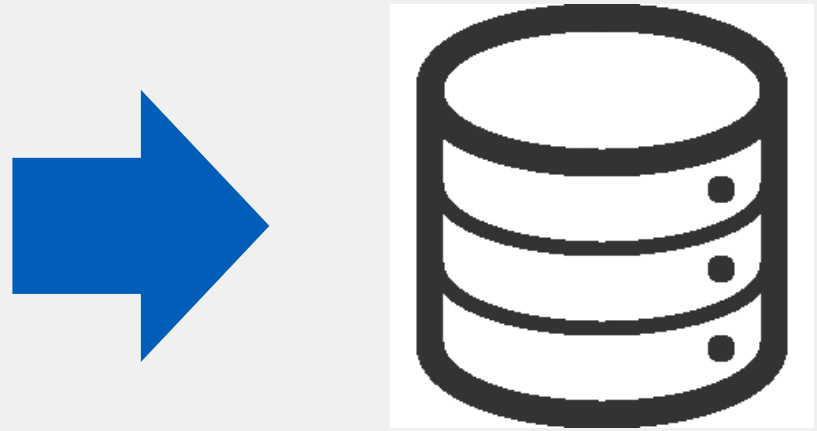
ARTIKEL 1. NAAM - RECHTSVORM
De vennootschap heeft de vorm van een besloten vennootschap.
Zij draagt de naam "***HENROSA***".

ARTIKEL 2. DUUR
De vennootschap is opgericht voor ***onbepaalde tijd***.
De vennootschap verkrijgt pas rechtspersoonlijkheid vanaf neerlegging van de uitgifte en het uittreksel van de oprichtingsakte op de griffie van de ondernemingsrechtbank waar de vennootschap haar zetel heeft conform artikel 2:7, §1 WVV.
De vennootschap kan ontbonden worden bij beslut van de algemene vergadering die beraadslaagt zoals inzake statutenwijziging.

ARTIKEL 3. ZETEL
De maatschappelijke zetel is gevestigd in het ***Vlaams Gewest***.
Het bestuursorgaan is bevoegd de zetel van de vennootschap binnen België te verplaatsen, voor zover die verplaatsing overeenkomstig de toepasselijke taalwetgeving niet verplicht tot een wijziging van de taal van de statuten. Dergelijke beslissing van het bestuursorgaan vereist geen statutenwijziging, tenzij de zetel verplaatst wordt naar een ander Gewest, in dit laatste geval is het bestuursorgaan bevoegd om tot de statutenwijziging te beslissen.
Indien ten gevolge van de verplaatsing van de zetel de taal van de statuten moet worden gewijzigd, kan enkel de algemene vergadering deze beslissing nemen met inachtneming van de vereisten voor statutenwijziging.
Iedere verandering van de zetel wordt door de zorgen van het bestuursorgaan in de bijlage tot het Belgisch Staatsblad bekendgemaakt.
De vennootschap mag, bij beslissing van het bestuursorgaan, exploitatiezetels, administratieve zetels, filialen, agentschappen en depots in België of het buitenland oprichten.

ARTIKEL 4. VOORWERP
De vennootschap heeft tot voorwerp :
- Het verlenen van adviezen binnen het financiële, aankoop en supply chain, duurzaamheid, HR en informaticadomein en begeleiden van de implementatie daarvan aan ondernemingen en de overheid op het gebied van planning, organisatie, efficiëntie en toezicht, informatie op het gebied van financieel beheer, begeleiding in veranderingstrajecten en projecten enzovoort.
Het voorzien van een gans gamma aan dagelijkse kantoordienstverlening, zoals de financiële planning, de facturatie, activiteiten met betrekking tot personeels- en informaticazaken, financiële transacties, de behandeling van post, de logistiek, enzovoort voor een vast bedrag of op contractbasis.
- Het aansturen en ondersteunen van een team medewerkers.
- Het geven van opleidingen en trainingen.
- Het organiseren van evenementen.

1



Business Value

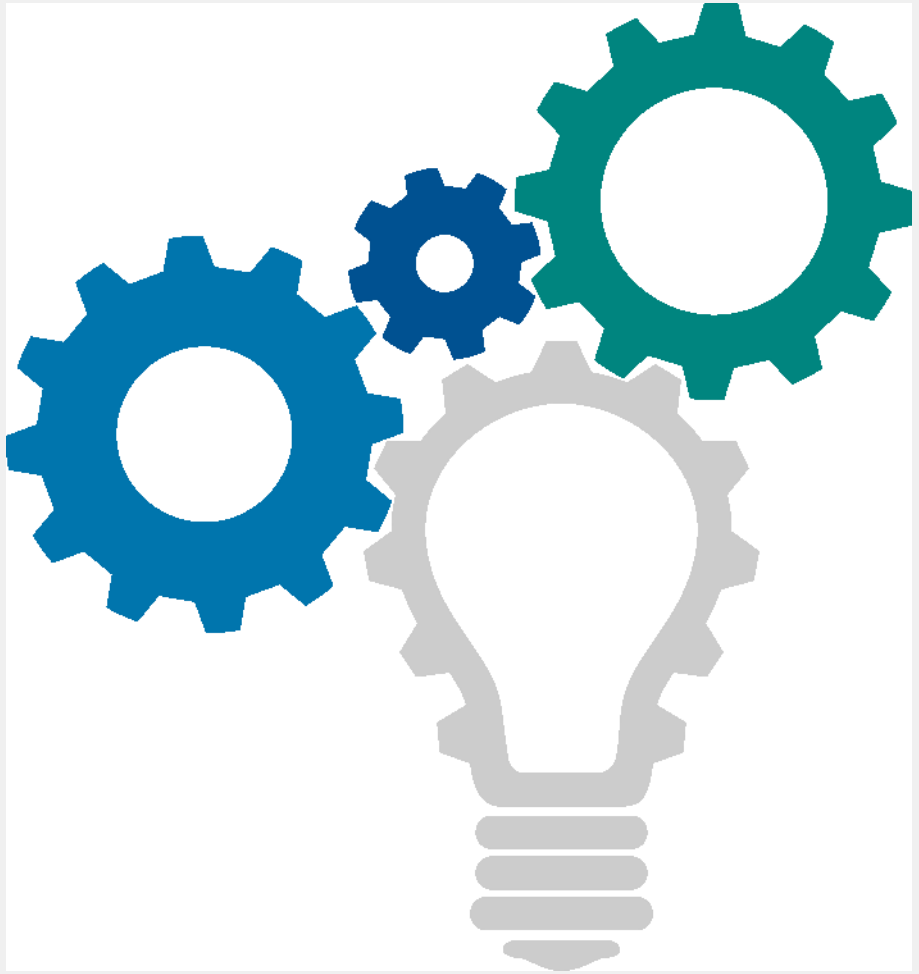
- **Permanent Database of Company Statutes**
  (instead of ad-hoc collection)

- **Automated solution for a certain business case**
  (instead of manual analysis of statutes)

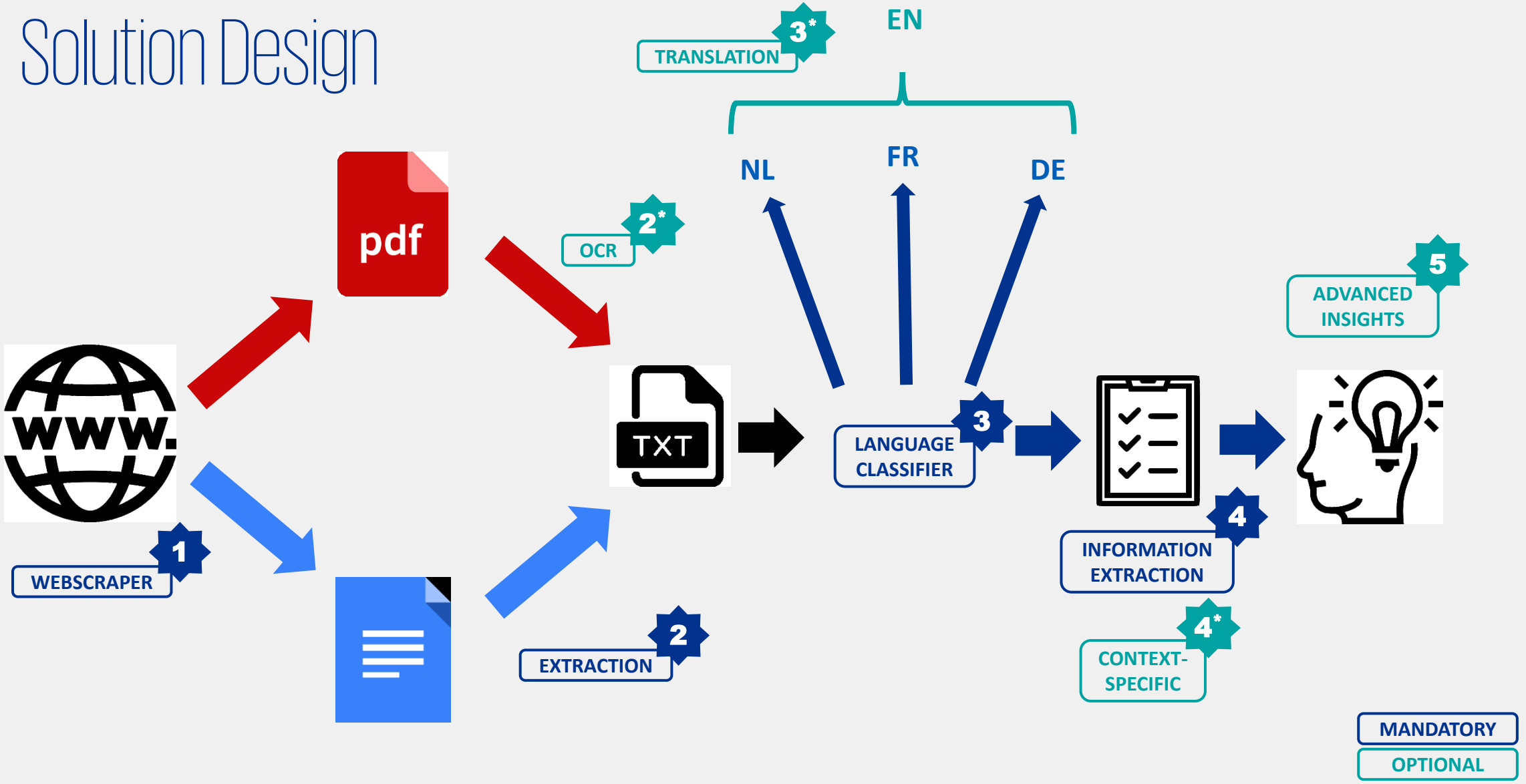
  → **Less time consuming**

  → **Higher capacity**

  → **Higher accuracy (?)**



Solution Design



1. Data Collection (Scraping)

Download statute pdf's

Aim for qualitative data: Textual > Scanned

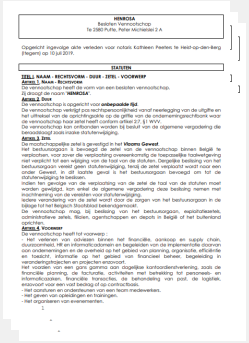
(optional) Add Meta Data when scraping (if possible)

Meta data as '<filename.pdf>.meta.json

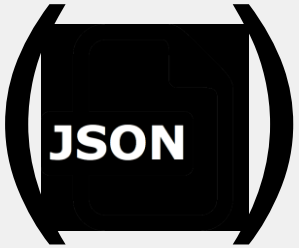
```
{
    "company_name": "…",
    "company_address": "…",
    "kbo_number": "…",
    "legal_form": "…",
    "associated_notary": "…"
}
```

TIPS

- **Interesting websites:**
  - High Quality Textual PDF Statutes: https://statuten.notaris.be
    JS-rendered, (but has an API)
  - Recently found companies: https://www.staatsbladmonitor.be/oprichtingen-bedrijven.html
- Don't overrun this website, work coördinated / divide workload, include 'sleep' steps during scraping, limit amount of downloads
  → **"scrape politely"**



123456.pdf



123456.meta.json

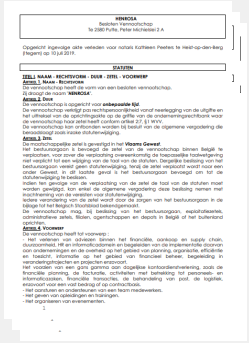
GOAL Construct sufficiently large & qualitative dataset

2. Text Extraction

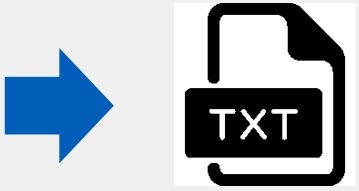
Extract the text from the pdf files.

TIPS

- **Keep ‘newline’ information;** the layout matters. (paragraphs, …)
- Useful python package**: pdfminer.six**



123456.pdf



123456.txt

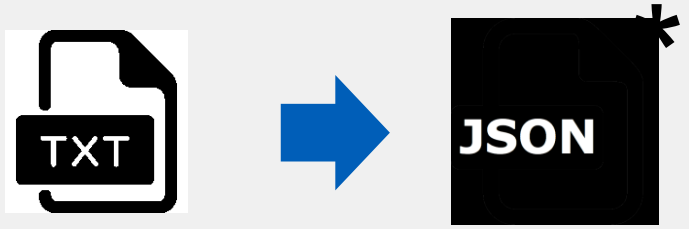
GOAL Transform pdf to plain text

3. Language Detection/Classification

For each document, detect the language and store the results.

TIPS

- **Don't reinvent the wheel:** Make use of webservice API's Azure, GCP, AWS) or python packages (e.g. Spacy, spacy-langdetect, …)
- **Append results to meta-data**



123456.txt ***123456.meta.json***

GOAL Classify documents per language

4. Unstructured Information Extraction

Mandatory:

- Articles (title + content)

Choose 2 (or more) from the list below:

- Company Name
- Legal Form (BV/SRL, NV/SA, VZW/ASBL, …)
- Company Address
- Associated Notary
- Date of Creation
- References (to articles / laws)

TIPS

- Focus on **one language (NL/FR)**
- **Start with the article detection**, then work context specific for the others
- **Combine both basic tools (like regex) and more advanced (statistical model)**

GOAL Add relevant metadata to the dataset

5. Advanced Insights

Some examples:

- **Classify** articles based on title & contents
- Search for **references** to (specific) laws
- Extract **legal-form specific** information
- **Compare** similar chapters with other companies in the same sector
- Extract the **total capital, number of shares**, …
- Check on the competences of company leaders
- Check whether company leaders are reimbursed
- Which formalities are required to gather the **General Assembly**?
- Extract the **relative date of the General Assembly** (e.g. 'First Monday of May')
- **…**

TIPS

- Focus on **one language (NL/FR)**
- **Start with the article detection**, then work context specific for the others
- **Combine both basic tools (like regex) and more advanced (statistical model)**
- Create manual article types or detect by clustering using with keywords

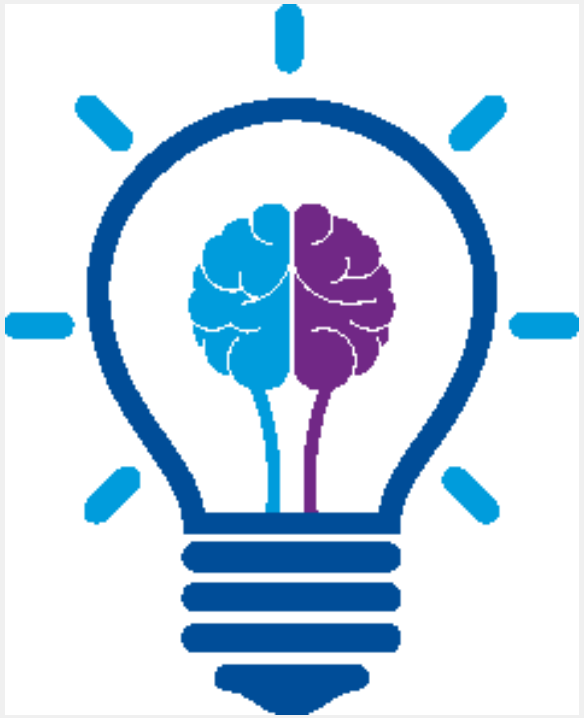
→ Be Creative!

GOAL Generate some interesting insights in the dataset

Expert Track

Each step can be expanded further:

1. Collect **Scanned PDF's** from KBO website
2. Extract text through **OCR** from scanned PDF's (Tesseract, Vision API, Azure, …)
3. **Translate** all documents to a common base language: (i.e. English)
4. Extract **context-specific** values (e.g. company funds amount from article)



TIPS

- First finish all the basic steps before attempting the expert track

General

TIPS & TRICKS

- Develop each ‘step’ as a separate component that works with an input file and produces an output file, ready for the next step. → pipeline
- Use a DocumentDB (TinyDB, Mongo DB) for storing your results. Or just use the file system.
- When training models, split test/validation set and measure accuracy/AUC, … (use meta-data as labels where available)
- Google search is your friend, but also don’t hesitate to share your questions in the Slack channel.

Practical

W1

- Data Collection
- Text Extraction
- **Deadline for collection: 13/09**
- Creating shared folder for the data

W2

- **Every group starts with same input**
- Language Detection
- Unstructured Information Extraction

W3

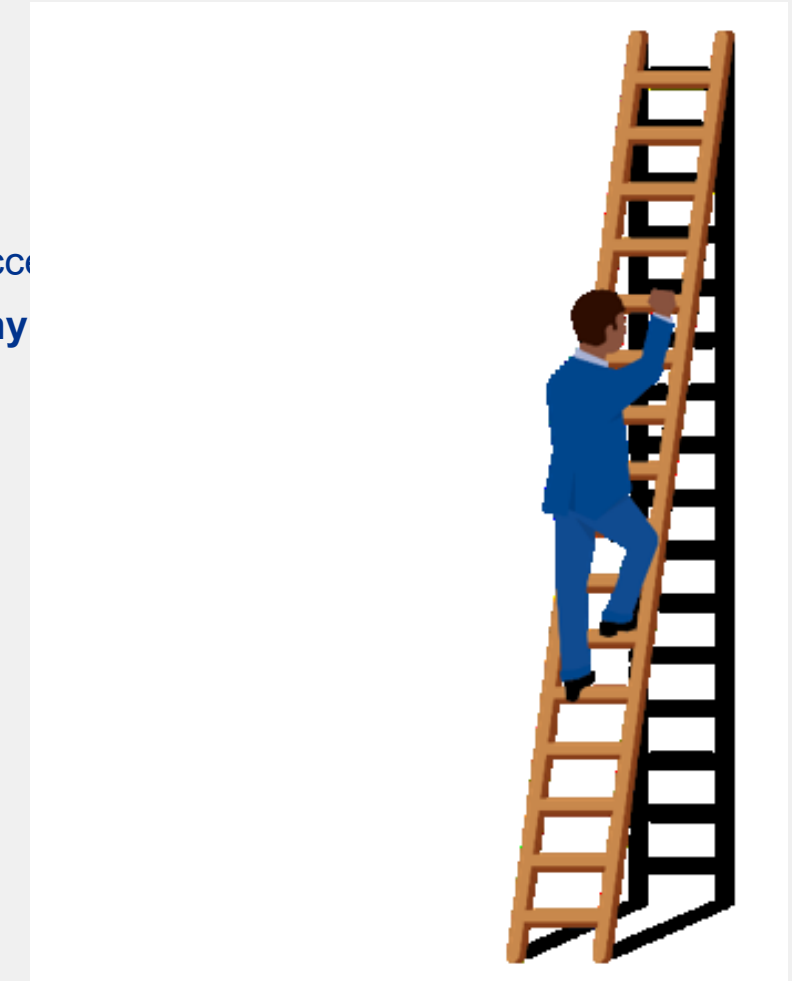
- Advanced Insights
- Present Findings (20 minutes)

Feedback

- The **internet** is your friend
- Don’t hesitate to send us your questions via **Slack**
- Avoid asking the same questions, **look around on Slack first**
- **Help each other**, even other groups
- We will regularly check Slack to answer your questions
- We will **come over** twice every week (time to be announced via Slack)
- During week 3, our feedback will include the **business side** as well (focus on your **presentations**)

Expectations

- Build **creative** solutions, think outside-the-box!
- **Collaborate** across teams where possible
- **Share** experiences and tips
- **Document** your approaches, where and why you failed / succ
- **Present** not only what you found, but **how** you did it, and **why**



Conclusion