Extraction of Public Ownership Data from Company Filings using Foreseer

February 2020

Overview

Foreseer is an enterprise-grade data processing platform that has the ability to source and extract unstructured data and let users enrich the extracted data for consumption. Foreseer is being used by a major global financial institution of repute to extract Ownership Data from public company filings across the world.

Use Case

Our client is interested in collecting company ownership data from public filings. The data typically comprises names of individuals and entities that hold substantial stake in companies and the quantum of their stake. Here is a sample of the Ownership data.

Fig 1. Ownership data reported in tables and footnotes

The Board and Auditors

Björn O. Nilsson

Chairman of the Board
Doctor of Science. Born 1956. Lives in Sollentuna, Sweden. Professor, CEO and member of the Royal Swedish Academy of Engineering Sciences during the period 2008–2017. Associate professor at the Royal Institute of Technology (KTH) in Stockholm. Member of the Board since 1999. Chairman of the Board since 2011. Chairman of the Remuneration Committee and member of the Audit Committee.

Other board appointments
Chairman of the Boards of the Swedish Foundation for Strategic Research, Swedish Young Academy Foundation, Swedish Athletic Association and Stockholm Science City. Member of the Boards of ÅF AB, European Institute of Innovation and Technology (EIT), SweTree Technologies AB and SwedNanoTech AB.

Shareholding
54,474

Vessela Alexieva

Employee representative
MSc in Molecular and Functional biology. Born 1959. Lives in Lund, Sweden. Senior Research Engineer. Member of the Board since 2013.

Other board appointments
-

Shareholding
20,850 (own and affiliated holdings)

Dharminder Chahal

M.Sc. in Aerospace Engineering and M.Sc. in Business Economics. Born 1976. Lives in the Netherlands. CEO of SkylineDx since 2013. He is also currently the CEO of Quorics, Managing Director at Exponential BV, and Fund Manager at Swanbridge Capital. Extensive board experience within life science in current and previous board roles at Agendia, Bioinvent (2013–2016), deVGen, Innate Pharma, and Octoplus. Member of the Board since 2017. Member of the Audit Committee.

Other board appointments
Chairman of the Board of DCPrime. Member of the Boards of Isobionics and VitalneXt.

Shareholding
250,000

Elin Jaensson Gyllenbäck

Employee representative
Ph.D. in Immunology. Born 1979. Lives in Lund, Sweden. Senior Research Scientist. Member of the Board since 2017.

Other board appointments
-

Shareholding
-

Fig 2. Ownership data reported in the last line of people biographies.

The data comes from SEC filings and unstructured PDF from international bodies, exchanges and companies themselves. The types of filings involve 8K, 10K, DEF 14-A, Annual Reports among others.

Value Proposition

The client wanted a solution to extract Ownership data from relevant public filings, enrich the extracted data by applying business rules the solution determines appropriate for the case from a large portfolio of business rules, and provide the final enriched data for the client to consume.

The client has been collecting Ownership data manually. Automating the data extraction and enrichment process was expected to:
- Deliver the client significant savings by means of reduced manual work in collecting data.
- Faster time-to-market, so they are able to publish their data to their customers faster than their competitors.
- Increased accuracy of the data collected since the solution is not susceptible to human error.

Challenges

Extracting data from unstructured documents is hardly a straightforward problem to solve. Here are some of the problems our engineering team at Foreseer had to solve to deliver a solution that helped our client extract unstructured Ownership data.

- **Identifying relevant data in large documents.** One of the preliminary problems that we had to solve was the ability for our solution to sift through large documents, eliminate noise and narrow down to a highly relevant set.
- **Recognizing table patterns.** Tables reporting Ownership data come in a variety of structures. The solution had to be capable of understanding all tabular structures.
- **Extracting data from running text.** The client wanted data reported in footnotes of tables and as well as elsewhere as running text to be identified and extracted. The extracted footnotes data are often adjusted with the tabular data using logic that the solution had to be made capable of implementing.
- **Applying business rules.** Rules prescribed by the client dictate how data is collected. Business rules help enrich data and get the extracted data closer to a form at which it becomes more directly consumable by the client.

Foreseer was able to solve these core problems in order to make the solution work for the client.

Solution

The Foreseer solution to extract unstructured Ownership data comprised the following components and features.

Fig 3. Foreseer Data Validation UI for users to review quality of the extracted data.

1. Client calls Foreseer API passing the document to be processed

2. Document passes through various Foreseer engines for the relevant data to be extracted

OCR Engine

Table Extraction Engine

Text Extraction Engine

3. Extracted data gets published in the Foreseer Validation UI for users to review data for accuracy

Business Rules

4. User reviewed extracted data gets relevant business rules applied on for the data to be enriched for user consumption

5. Final output gets published in the Foreseer Validation UI for users to review if they prefer

6. Final output gets sent to the client database via the Foreseer API
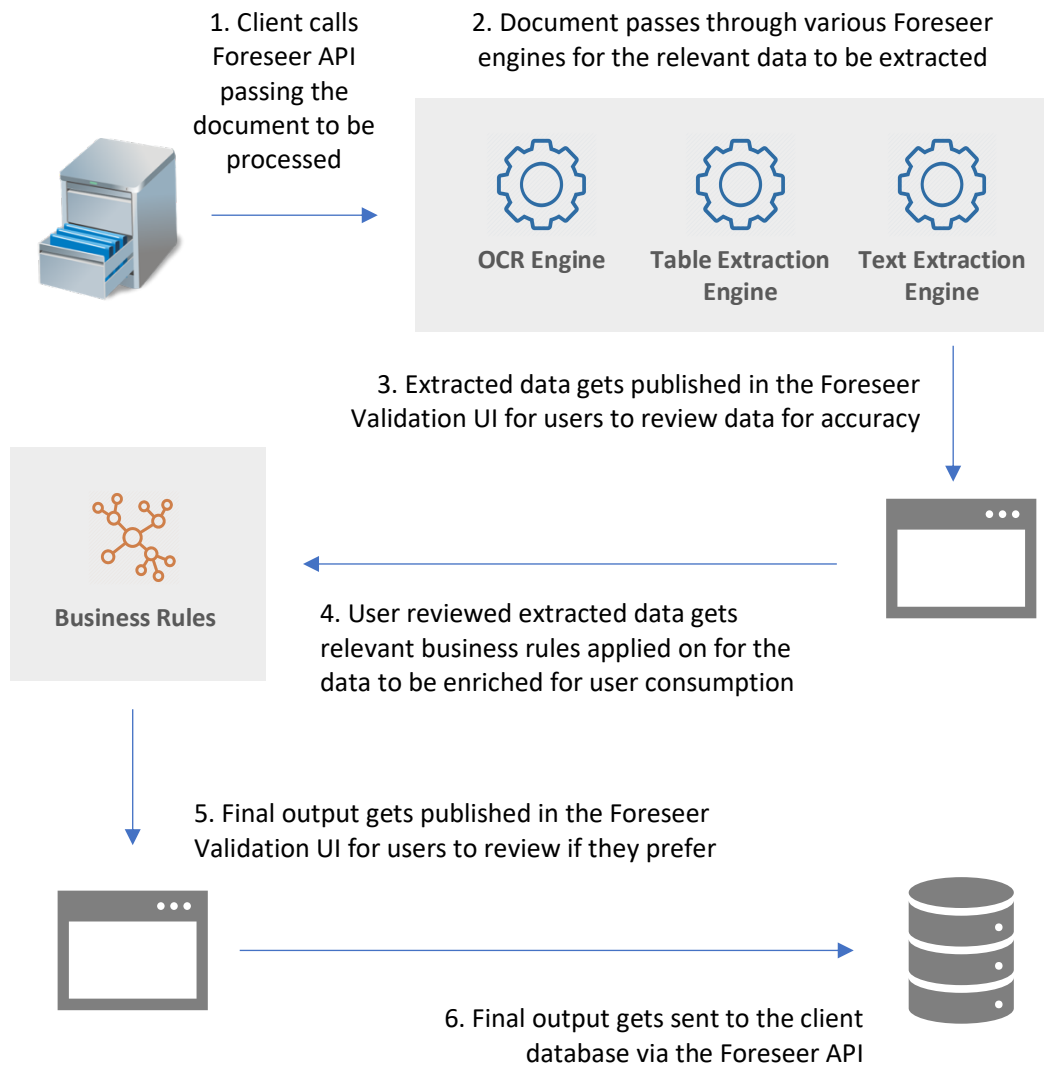
Fig 4. Architecture of the solution deployed for the client.

- **Data Extraction Engines:** Built using supervised machine learning, capable of identifying and extracting relevant data from unstructured documents.
- **Data Enrichment Framework:** Providing users the ability to create and maintain data collection/enrichment business rules that run on the extracted intermediate data to deliver the final output.
- **User Interface:** To help users vet the quality of the extracted and enriched data and make changes where essential. User changes are logged and fed back into the models for continuous learning.

- **API:** To facilitate submission of documents by the client, with optional callbacks to be notified at critical checkpoints in the processing pipeline.
- **Reduced Source Document:** That is a condensed version of the original document with pages reporting irrelevant data removed so users are able to get to the information that they are interested faster.
- **OCR Engines:** To convert scanned documents to text-based extractable documents for the data extraction models to operate on.
- **Downstream Publication:** Once the data reaches its final signed-off state, it is sent to the client in the requested location and format.

Conclusion

Using Foreseer's solution, the client was able to extract Ownership data from unstructured documents with little manual intervention, thus reaping significant efficiency gains, faster go-to-market and increased quality of published data.

About Foreseer

Foreseer is an enterprise-grade data extraction platform that enables businesses to extract information from unstructured documents, complex PDF, scanned documents and social media.

Foreseer offers complete human-in-the-loop solutions along with data labeling services if you need to augment your data preparation workforce. Foreseer is an open platform that your data scientists and business users can leverage to unlock value without worrying about data, technology and infrastructure heavy-lifting necessary to deliver a scalable, robust, enterprise-grade solution.

If you want to learn more about Foreseer, visit our website at www.foreseer.ai or email us at hello@foreseer.ai.