DIGITALISATION CASE STUDY

AUTOMATED DATA ENTRY

FOR

FINNISH COMMERCIAL BANK

DIGISALIX

www.digisalix.fi

80% FASTER
DATA EXTRACTION
FROM
BUILDING MANAGER'S CERTIFICATES

DIGISALIX

Summary

We reduced manual typing required to serve mortgage and real estate customers up to 80%. That means more time for the customer encounter, less for typing.

Our customer, The Bank, only had to provide us a modest amount of example documents and an access to their private cloud for us to install a custom smart scanning system for “building manager’s certificates” in 8 short weeks.

SMART DOCUMENT SCANNING IMPROVES THE BANK’S CUSTOMER SERVICE

- 80% less typing means more time to serve the customers.
- 12 minutes time savings per customer/certificate.
- Document data is now instantly structured into a database-friendly form.

Even complex data types were extracted, e.g. lists of building renovations.

The Digisalix machine learning system was trained with only 848 documents. Instead of using big data, we gave structural hints for the computer to understand the content. This small data approach enables efficient setup without compromising the accuracy of the system.

DIGISALIX

1. Challenge & solution overview

We automate data entry with our smart document scanning system that operates in the customer’s private cloud. When the incoming documents are automatically processed, people can stay focused on their actual, productive work instead of typing in the data.

DIGISALIX

AUTOMATION ROADBLOCK

The data entry problem

Data entry is one of the most redundant and hated office tasks, but necessary in order to get computers to process the data they can't read from documents on their own. We are changing this.

Professionals from many walks of life fall victims to data entry. Roughly a fifth of a nurse's workday is spent on it and there is a $4B industry in the US focused solely on this mindless typing. Hidden data entry tasks that occur in many office jobs is the real problem, and also bank clerks suffer from them.

Automating this nuisance is not as trivial as one might think – therefore it's still mostly done manually. We are changing this by teaching computers to read better.

Our goal is to automatically process the incoming documents, so people can focus on their actual, productive work – in this case – serving the mortgage and real estate customers of The Bank.

By automating we can also reduce delays and errors that often occur in repetitive tasks performed by humans.

THE AUTOMATION NEED FOR THE

Data Entry of Building Manager's Certificates (BMC)

For each mortgage and real estate customer of the Bank, the clerk uses 15 minutes on typing in the data from BMC. Annually this means thousands of hours that could be used better. Let's automate.

The Building Manager's Certificate is required when apartments are sold or mortgage applications are evaluated. BMC describes many technical details of an apartment, starting from the size and configuration all the way to performed and planned renovations in the building. The BMC's come in myriad of layouts, often spanning 3–5 pages and including free-form fields with no predefined information structure. Unfortunately, computers rely on structure to process information.

The aforementioned intricacies pose a complicated automation task, which nonetheless is attainable with Digisalix's combination of machine learning magic, clever coding and careful training of the system.

CUSTOMER FRIENDLY SMALL DATA APPROACH

What does it take to get started?

Our small data machine learning solution is fast and light in many aspects. Starting the automation process is easy for the customer – The Bank only needed to provide us:

- Access to set up the extraction system to a private cloud where the confidential data was securely managed by The Bank's IT security.
- 848 building manager's certificates to configure and test the data extraction system for the certificates. This is small data – more on that later.
- The specification of the 19 fields to extract from the certificates.
- A contact for questions about ambiguous fields.

With these ingredients and Digisalix's toolkit, our team independently set up, trained and integrated the extraction system in only 8 weeks. The progress was demonstrated in biweekly meetings.

LET'S GET STARTED

Deployment and usage

To automate the certificates, Digisalix installed a server software into the bank's private cloud. The server receives PDFs from business applications and responds with values of the fields. These values are now in a structured form and therefore simple to process in the business applications and to store into databases.

Automated data entry for business applications can be put to service in different ways:

- A customer service agent uploads the certificate and fields of interest are automatically picked up. The agent can quickly glance over the results.
- Certificates for, e.g., online loan applications are processed separately before evaluating the applications. Specialists perform quality control over the results.
- Archives of certificates transformed into structured database enable searches, gathering business intelligence, and modelling real estate markets.

The next section, Results, covers the accuracy and time savings from the reduced typing.

2. Results

We improved the speed of data entry significantly (estimated 80%) and reached high accuracy for extracting the data (up to 99%). Our system is also built to highlights cases where human review is advised to reach the best possible result.

DIGISALIX

Extraction accuracy

Reliability of the produced data is essential. To evaluate the performance of the system, we checked how often it extracted the correct information for each of the 19 fields in a separate test set* of documents.

“Key–value” fields

Simple key–value fields achieved accuracy up to 99%, average being 93%. Key–value pairs often have a small set of keywords in vicinity of the actual desired value.

Sometimes even the simple values have room for interpretation, for example, classification of various monthly charges to capital expenditures and maintenance costs.

UP TO

99%

EXTRACTION ACCURACY

AVERAGE

93%

EXTRACTION ACCURACY

*The full set of 848 documents was randomly split into non-overlapping training (630) and testing (218) sets.

DIGISALIX

Extraction accuracy

Complex lists

More complex fields — such as renovation lists – were the interesting part. For example, each renovation list item consists, ideally, of a year and a description. They might be presented as comma-separated free-form text, clean tables with columns, or anything in between.

These kind of lists are the time consuming part of data entry, and also the hardest to automate. **Computers need structure** to ”understand” the data, and **humans type with plethora of randomness** if they are not strictly guided.

Renovation lists were correctly read in 51% of the test documents, which might not seem much, but it does reduce the workload significantly. Vast majority of the remaining 49% had only minor errors that were corrected with minimal effort from the user. 93% of all the entries across the documents were found successfully.

93%

ITEMS FOUND

51%

RENOVATION LISTS HANDLED CORRECTLY

Improved performance

Even though the contemporary data extraction systems can’t be expected to reach 100% automation rate, they can still make a huge difference by changing the mindless data entry work to mostly supervising the results. The supervision is only needed when the system is not confident about the extraction outcome.

As the system becomes generally adopted, we expect to save 80% of time spent by the bank's clerks on data entry of Building Managers Certificates. Annually thousands of hours can be focused on more productive work.

Time savings will be reached as a combination of full automation and typing converted to checking results. No one has to learn new skills or tools, since Digisalix system integrates to the familiar tools The Bank is already using.

80%

FASTER OVERALL
DATA EXTRACTION

Impressive, isn’t it? Contact Kalle to find out how we can help you, or keep on reading!

GET IN TOUCH

DIGISALIX

3. Solution

We teach computers to read by “understanding the structure”. No need for big data, neural networks, or hard to maintain templates. We do it with small data, combining machine learned components and domain knowledge in an adaptive inference engine.

DIGISALIX

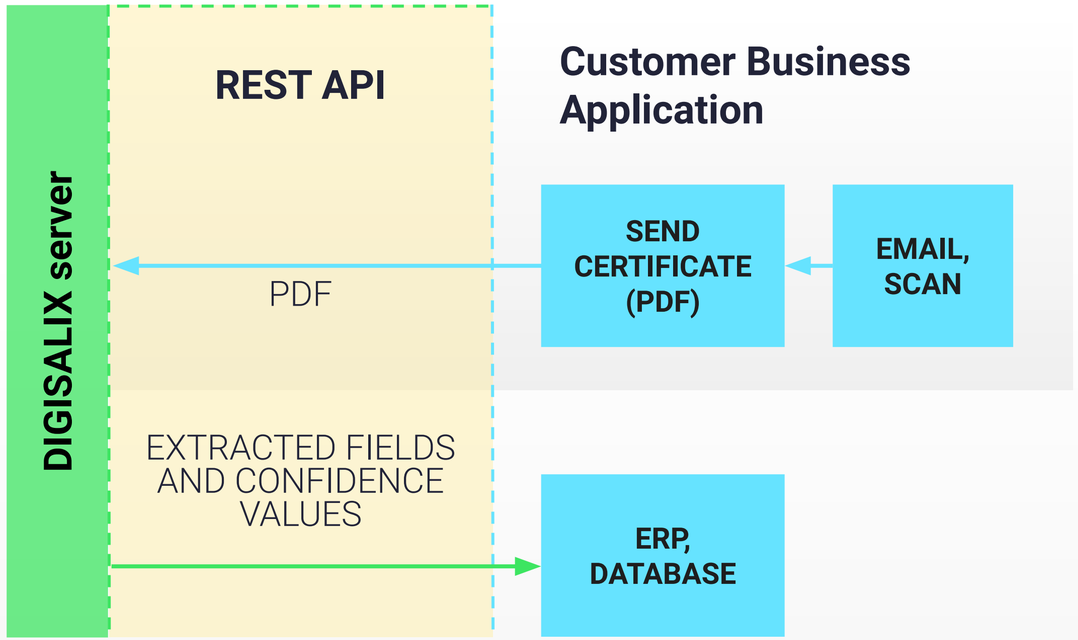
Smart scanning solution for building manager's certificates

Digisalix smart scanning solution is a REST API Server that interprets documents – in this case the building managers certificates – into a standard structured format. Any business application can easily automate the processing of the certificates as the Digisalix server converts hard-to-process certificates into clean structured data.

As the data extraction and business application are split, the customer can independently implement and improve workflows. Digisalix takes care of the data extraction.

In this section you will learn:

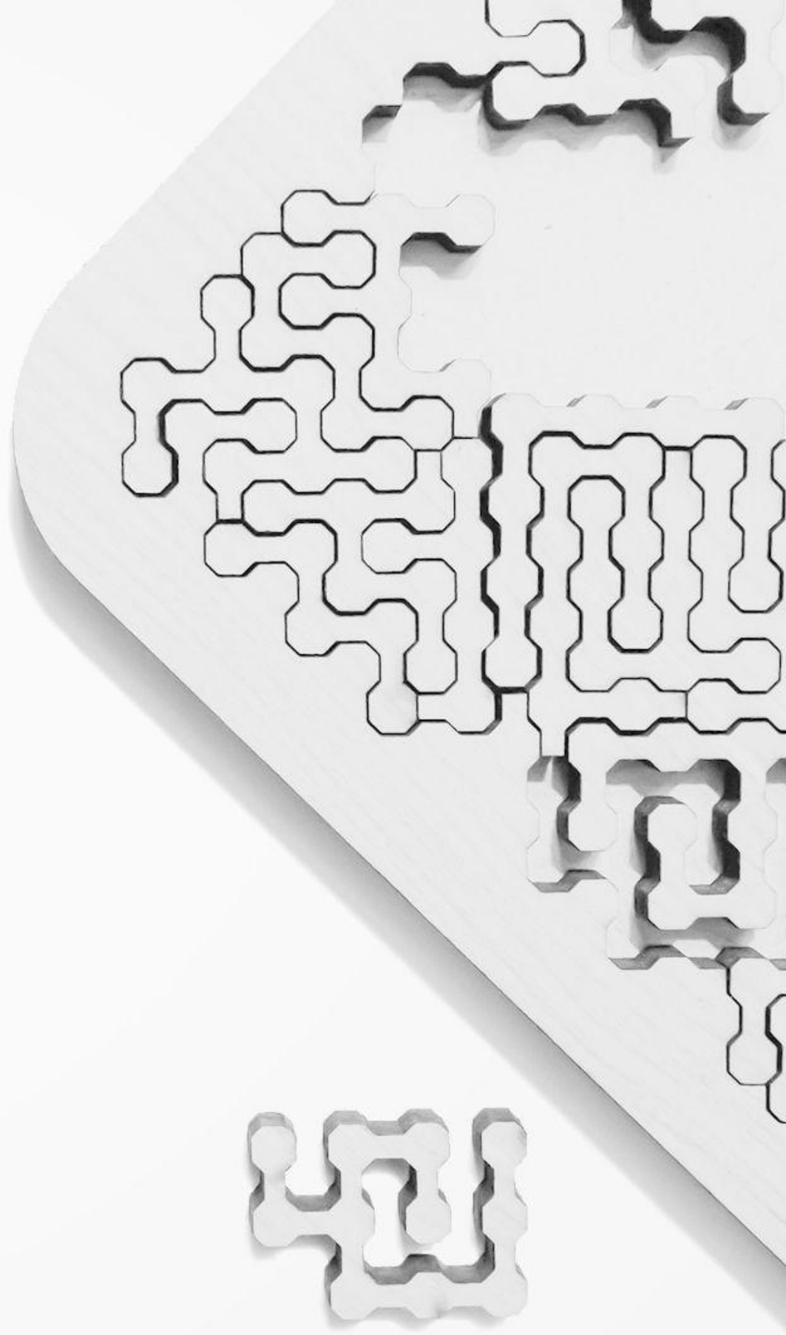
- Purpose and meaning of structure (and life)
- How Digisalix sets up the system for a new task, the certificates
- How to use system reported confidence to optimise automation rate



THE DIGISALIX SOLUTION

Computers need structure to “understand” the data

Think of document automation as a jigsaw puzzle. Computers don't inherently possess a model of the world, so they cannot assemble the pieces by purely looking at the raw data. However, after a few key pieces are provided, the structure of the picture starts to reveal and the puzzle becomes markedly easier to solve. The computer, with a small amount of data, can now fill in the rest.



Structural definition and small data

Our software makes it easy to provide the crucial key pieces in the form of a structural definition of the document. Now the system can complete the picture via machine learning, using small, well-curated datasets that demand little human labour and maintenance to keep at high quality. Light annotation loads are adaptive: adding new fields or changing existing ones takes only some hours.

When all pieces fit snugly together, we can be confident about the final result. And if something seems amiss, it's easier to analyse the issues compared to approaches that use highly flexible machine learning models that subsume everything into one huge parameter space. Our approach is compatible with such models too, but we prefer to use them as components for specific tasks when lighter approaches are not enough.

Setting up the system

1. Defined the fields and their properties using the Python API of the Digisalix Engine.

2. Annotated certificates for the defined fields.

3. Trained machine learning (ML) models for locating each field.

4. Analysed extraction accuracy.

The workflow was iterative and annotations were done in batches. As more documents were used for training, the ML models improved and the process became faster. The field definitions improved as we learned about the contents of the certificates.

Sometimes learnings lead to changing existing annotations. As there were only some hundreds of documents involved, annotations of one field could be reviewed and updated quickly.

DIGISALIX

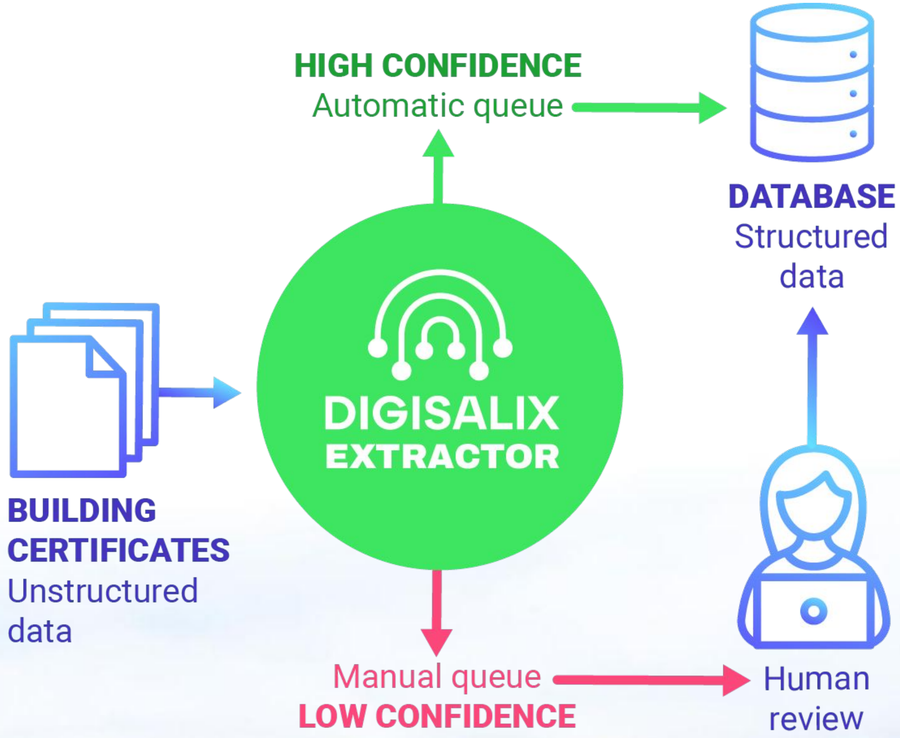
Automatic confidence estimation

The data extraction can be problematic if the certificate is poorly scanned, the author of the certificate has been creative with formatting or the data is ambiguous. To discern between easy and hard cases the system estimates how certain it is of the results. The estimate is called confidence.

Based on the confidence level, humans are notified to review uncertain cases, while others can be trusted to be correct and go straight to the database.

So, extracted documents are divided into **automatic** and **manual** queues based on predetermined rules for the confidence and importance levels.

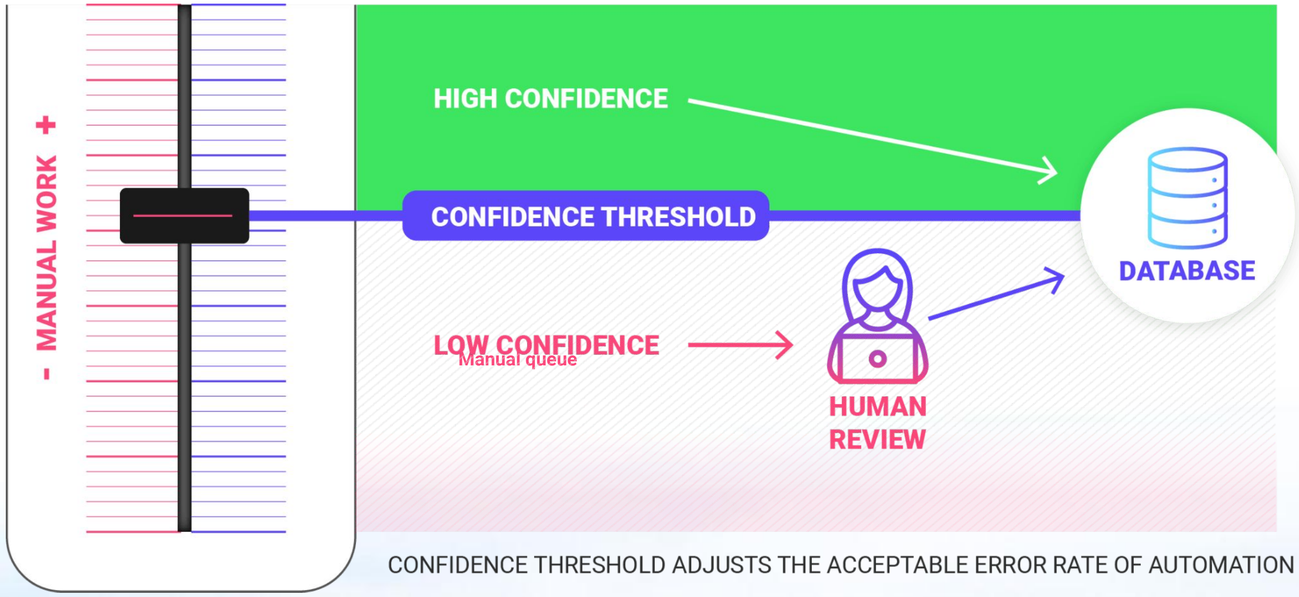
Extracted results help the manual queue review significantly, making it more about checking the results than typing in the values.



Confidence threshold balances automation rate and accuracy

Customer can adjust the automation rate **(CONFIDENCE THRESHOLD)** for the whole document or single fields, based on their business requirements and processes.

- The cost of errors in **automatic queue** vs. unnecessary reviews in **manual queue** are generally different.
- The threshold can be freely set by the business application to find the best balance. Digisalix will help in finding good thresholds. We also continuously optimise the values as part of normal maintenance.



CONFIDENCE THRESHOLD ADJUSTS THE ACCEPTABLE ERROR RATE OF AUTOMATION

DIGISALIX

Small data approach to machine learning

Big data combined with deep learning has shown good results in, e.g., image analysis, recommendation systems, and text synthesis. However, utilising big data has its limitations, and as AI-superstar Andrew Ng said: “In many industries where giant data sets simply don’t exist, **the focus has to shift** from big data to **good data.**”

Thanks to Small data approach, Digisalix can

- solve use cases backed by limited data,
- tailor to customer needs instead of one-size-fits-all model,
- quickly re-train models with less compute resources spent,
- ensure the training data is well curated and models documents correctly, and
- deploy the system without GPUs or other expensive compute elements.

When we combine several simpler, small data machine learned components, we create rich behavior but with much lower costs than complex models built on big data.

SMALL DATA =

-Energy & time efficient training, with fast iteration cycles

DIGISALIX

4. Technical details

The Digisalix Doclib Engine forms the core of our software architecture, providing easy integration into customer's business applications and fast customization to different document types. Our Review GUI web app connects directly to Doclib for integrated annotation workflow.

How the system works?

We teach computers to read by “understanding the structure”. No need for big data, neural networks or hard to maintain templates. We do it with small data, combining machine learned components and domain knowledge in an adaptive inference engine.

To extract structured data from document images, the system

1. retrieves the text snippets and their locations using optical character recognition,
2. locates the fields of interest based on logical and structural cues on the content and relationships of the fields,
3. parses and normalizes the texts into final structured values (e.g., discard unnecessary text and convert dates into ISO format),
4. estimates the confidence scores for the values.

The steps are performed with our Doclib Engine, configured for the specific extraction task using the Python API of the Engine. The configuration describes the fields, their constraints, and dependencies – the structure of the information. Combining data and structure into a self-correcting inference engine helps the system reach high accuracy with small data.

See the architecture diagram on the next page for additional details.

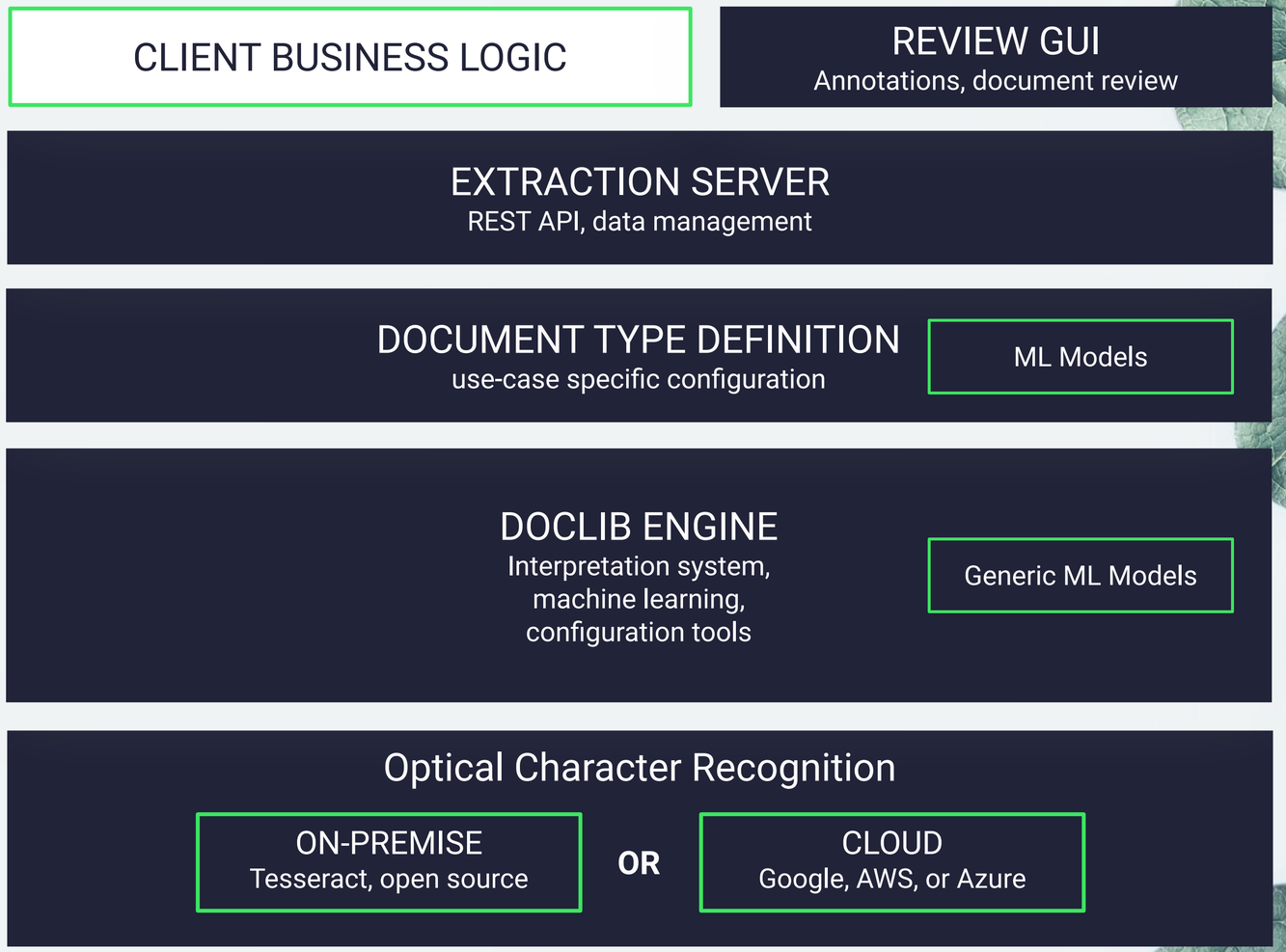
Architecture

Server for REST API and data storage. Business logic and GUI for annotations both use the same API.

Document type specific structure definition and field specific models.

Doclib: Machine learning facilities and general tools and models (e.g. key-value analysis).

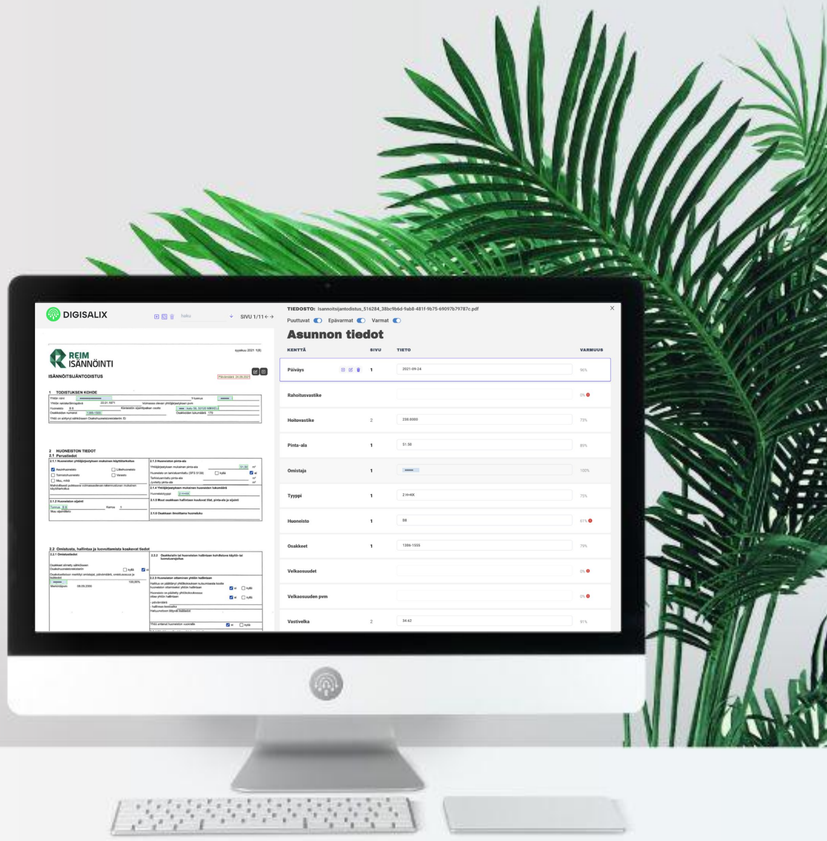
Optical character recognition provides the raw text and layout. Any OCR can be used.



DIGISALIX

ANNOTATIONS WERE DONE IN THE

Digisalix Review GUI tool



Modern web-browser-based application for annotation and review of extracted documents

- Documents securely on the server.
- Only HTTP/REST connection needed.
- Connects to the Digisalix extraction server

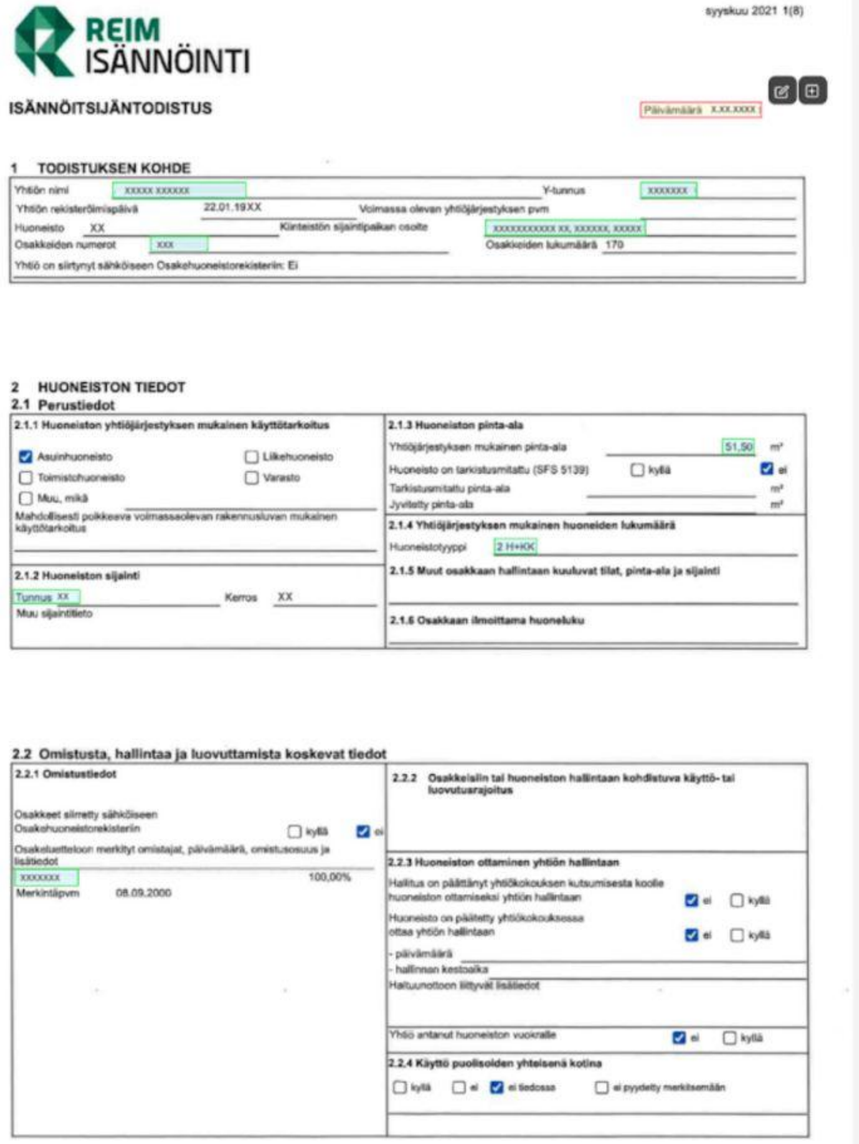
DIGISALIX

Review GUI

DIGISALIX

haku

SIVU 1/11



REIM ISÄNNÖINTI

syyskuu 2021 1(8)

ISÄNNÖITSIJÄNTODISTUS

Päivämäärä x.xx.xxxx

1 TODISTUKSEN KOHDE

Yhtiön nimi xxxxx xxxxxx — Y-tunnus xxxxxxx

Yhtiön rekisteröimispäivä 22.01.19XX — Voimassa olevan yhtiöjärjestyksen pvm

Huoneisto XX — Kiinteistön sijaintipaikan osoite xxxxxxxxxx xx, xxxxxx, xxxxx

Osakkeiden numerot xxx — Osakkeiden lukumäärä 170

Yhtiö on siirtynyt sähköiseen Osakehuoneistorekisteriin: Ei

2 HUONEISTON TIEDOT

2.1 Perustiedot

2.1.1 Huoneiston yhtiöjärjestyksen mukainen käyttötarkoitus

☑ Asuinhuoneisto ☐ Liikehuoneisto

☐ Toimistohuoneisto ☐ Varasto

☐ Muu, mikä

Mahdollisesti poikkeava voimassaolevan rakennusluvan mukainen käyttötarkoitus

2.1.2 Huoneiston sijainti

Tunnus xx — Kerros XX

Muu sijaintitieto

2.1.3 Huoneiston pinta-ala

Yhtiöjärjestyksen mukainen pinta-ala 51,50 m²

Huoneisto on tarkistusmitattu (SFS 5139) ☐ kyllä ☑ ei

Tarkistusmitattu pinta-ala m²

Jyvitetty pinta-ala m²

2.1.4 Yhtiöjärjestyksen mukainen huoneiden lukumäärä

Huoneistotyyppi 2 H+KK

2.1.5 Muut osakkaan hallintaan kuuluvat tilat, pinta-ala ja sijainti

2.1.6 Osakkaan ilmoittama huoneluku

2.2 Omistusta, hallintaa ja luovuttamista koskevat tiedot

2.2.1 Omistustiedot

Osakkeet siirretty sähköiseen Osakehuoneistorekisteriin ☐ kyllä ☑ ei

Osakeluetteloon merkityt omistajat, päivämäärä, omistusosuus ja lisätiedot

xxxxxxx 100,00%

Merkintäpvm 08.09.2000

2.2.2 Osakkeisiin tai huoneiston hallintaan kohdistuva käyttö- tai luovutusrajoitus

2.2.3 Huoneiston ottaminen yhtiön hallintaan

Hallitus on päättänyt yhtiökokouksen kutsumisesta koolle huoneiston ottamiseksi yhtiön hallintaan ☑ ei ☐ kyllä

Huoneisto on päätetty yhtiökokouksessa ottaa yhtiön hallintaan ☑ ei ☐ kyllä

- päivämäärä

- hallinnan kestoaika

Haltuunottoon liittyvät lisätiedot

Yhtiö antanut huoneiston vuokralle ☑ ei ☐ kyllä

2.2.4 Käyttö puolisoiden yhteisenä kotina

☐ kyllä ☐ ei ☑ ei tiedossa ☐ ei pyydetty merkitsemään

TIEDOSTO: Isannoitsijantodistus_516284_38bc9b6d-9ab8-481f-9b75-69097b79787c.pdf

Puuttuvat Epävarmat Varmat

Asunnon tiedot

KENTTÄ	SIVU	TIETO	VARMUUS
Päiväys	1	2021-09-24	96%
Rahoitusvastike			0%
Hoitovastike	2	238.0000	73%
Pinta-ala	1	51.50	89%
Omistaja	1	XXXXXX XXXXXX	100%
Tyyppi	1	2 H+KK	75%
Huoneisto	1	XX	61%
Osakkeet	1	XXXX – XXXX	79%
Velkaosuudet			0%
Velkaosuuden pvm			0%
Vastivelka	2	34.62	91%

Yhtiön tiedot

KENTTÄ	SIVU	TIETO	VARMUUS

5. Conclusion

We have presented a practical system that reduces typing work >80% for a complex form with varying layouts and complex entities such as lists with internal structure.

Case in numbers

848 building certificates annotated (typically 3–5 pages long)

19 fields ranging in difficulty between certificate date and renovation list

750 documents annotated by humans in 16 days interleaved with technical work.

10–20 documents per hour annotation speed.

Up to

99%

EXTRACTION ACCURACY

80%

FASTER OVERALL DATA EXTRACTION

Solution benefits

PRIVATE & SECURE

System can be set up in customer's private cloud – Data is under customer's control at all times.

SMALL DATA APPROACH

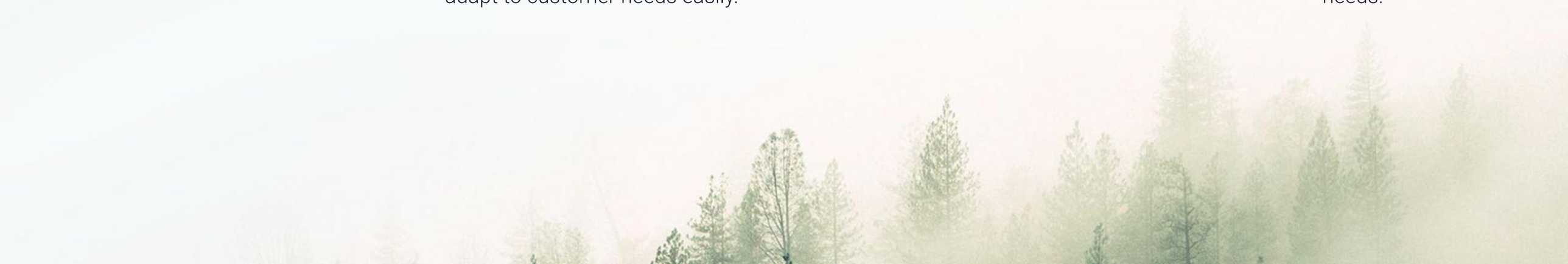
Faster development cycles, simplified maintenance, reduced analysis work and machine learning computing costs. Thanks to small data we can adapt to customer needs easily.

EXTENSIBLE TOOLKIT

Customers and partners can implement solutions independently, thanks to document types defined in Python code. Hard parts, like machine learning are handled by the Engine.

WORRY FREE

A turn-key full service solution: Annotation and model maintenance handled by Digisalix, based on customer needs.



Let’s do less typing!

Contact Kalle to discuss your document automation dreams. He’s happy to help!

Kalle Raita

kalle@digisalix.fi

Twitter LinkedIn

www.digisalix.fi

DIGISALIX