An AI-Based Approach to the IRS Form 990

Case Study

AYR.ai

Client
A leading science and technology provider

Problem
Current vendors in the market can only achieve 79% accuracy

Result
99%+ accuracy using our SingularityAI platform

Client

Client is a leader in science and technology solutions and works closely with the government and commercial businesses. They have years of experience working with tax practitioners and third parties doing business electronically with the IRS. Their expertise in software programming and development help enhance system functionality, increase system up-time, and reduce costs for their customer.

Problem

The IRS Form 990 is an informational tax form that most tax-exempt organizations must file annually. The form gives the IRS an overview of the organization's activities, governance and detailed financial information, and is used to prevent organizations from abusing their tax-exempt status.There are hundreds of thousands of Form 990s submitted yearly, each containing ~12 pages and 1000+ fields.

At first glance, the Form 990 may appear to be a simple, “structured” document (the document format/layout/appearance remains consistent); however, upon closer evaluation, one can say that the Form 990 possesses "unstructured" elements. For example, when using an electronic form filler, a table cell may expand to accommodate longer responses. In this instance, the Form 990 deviates from the original document template and can be considered a variation.

Current vendors in the marketplace can only achieve 79% accuracy. Additional time must be spent capturing the remaining 21% of values. Our client wants a scalable solution that can accurately extract all data from the Form 990 with high accuracy, regardless of document format, structure and variability.

AYR.ai

Challenges We Faced

Information Density

The IRS Form 990 contains many fields (1000+) across twelve pages that need to be extracted, which makes training a model that can accurately detect and extract each value more difficult

Uneven Distribution of Fields

Some fields may be left empty by the individual filling out the form. To train a robust model to recognize all potential values, however, we need training data that reflects all possible outcomes

Accuracy

The IRS Form 990 is a federal tax reporting requirement - accuracy is of the utmost important for auditing purposes

Scalability & generalization

Our solution should be able to handle document variation (e.g., different cell sizes, paper size, etc.), as well as be applied to similar-looking documents such as other forms within the Form 990 family

Checkboxes

Checkbox detection is challenging due to the relatively small size of checkbox markings, distinguishing between different marks (a checkmark vs an “x”), and distinguishing between a mark vs “noise”

Our Process

Intelligent Data Simulator

Using AYR's Intelligent Data Simulator (IDS), we generated synthetic, augmented data with a wide range of values to ensure our model can recognize every field. The IDS also automatically labels *all* the generated fields so there is no need to manually label them. This tool enables us to quickly build our models and scale our project capabilities.

Table Structurization

For speed, scalability, and flexibility reasons, we decided to train a table detection model for pages in the Form 990 containing fillable tables.

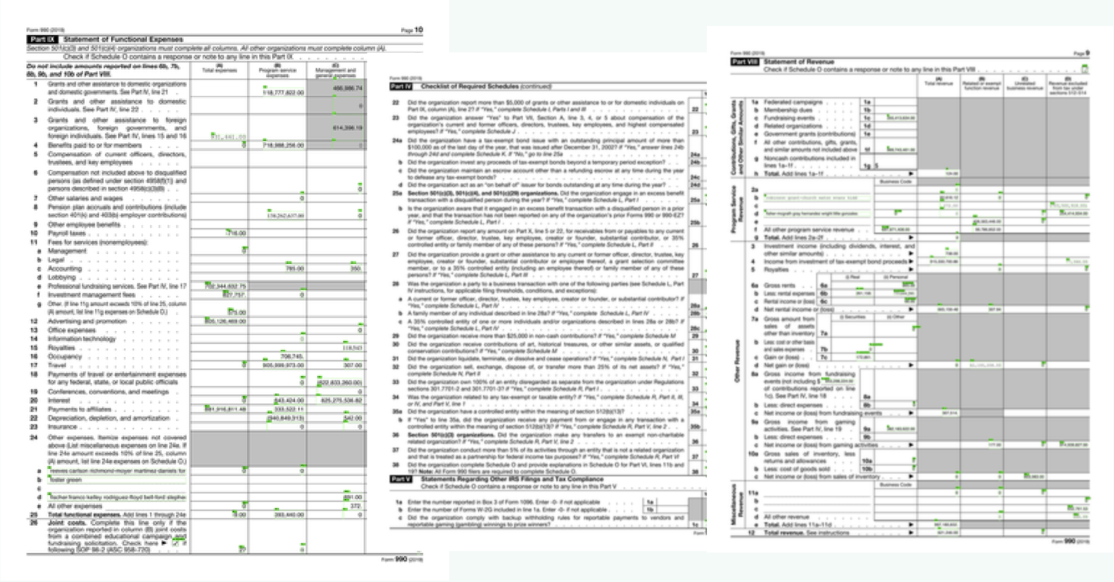
By performing table structurization, we no longer need to pre-define the schema belonging to the table (the schema in these tables are defined during post-processing). This means if changes are made to the table (e.g. a new row is added), it would not affect our model’s ability to make predictions. Furthermore, structuring data in this way enables us to accurately identify key-value pairs, as well as preserve the relationships between different values (e.g. beginning-of-year and end-of-year amounts).

Checkbox Detection

Several sections within the form contain checkboxes. By generating synthetic data for different check marks and different combinations of selections, we were able to train a model to accurately detect the checked field regardless of the symbol in addition to the text next to it (if applicable).

Our Solution

After using our SingularityAI platform to train a model, the customer can simply upload an image of the Form 990 and effortlessly extract all the values defined in the project. These values can be exported in multiple formats (e.g. HTML, JSON, etc) and integrated into downstream systems.



Sample Form 990 pages with our detection model (Pages 10, 9 and 4 from left-to-right)

Results

- 99.8% accuracy
- 21X reduction in human effort required (less than 1% of the forms submitted now require human review vs. previous 21% of forms)
- $20M+ potential annual cost savings across all processed IRS forms

About AYR

With its world-class team of scientists and developers, AYR has pioneered new AI techniques that have modernized and democratized Intelligent Document Processing (IDP). The company provides SingularityAI, an Artificial Intelligence platform enabling enterprises to transform their raw data into actionable insight.

Enterprise leaders use SingularityAI to efficiently convert high-volume unstructured content into machine-readable data, enabling real-time decision-making and powering improvements in customer experience and operational agility.