Don't Leave 80% - 90% ROI on the Table

Utilizing Next Gen Tech to Analyze & Extract Unstructured Data

For years, enterprises have been leveraging technology to analyze and extract structured data. But with only 10% to 20% of the total data within the enterprise being structured data, what happens to the remaining 80% to 90% of unstructured data? Frank Casale, IRPA AI Founder, sat down with Rohail Khan, Digital Exchange Solutions President, to discuss how applying automation to unstructured data can increase ROI by reducing processing times, the key challenges related to unstructured data management and how next gen technologies are yielding significant benefits to enterprises in this exclusive 3x1 (3 questions, 1 expert) interview.

Q There is significant amount of interest and activity in the area of addressing unstructured data. Why so ?

A The industry's been developing tools for making sense of and speeding up the extraction of information and intelligence available within structured data for decades. Most of our databases types, reporting engines, big data architecture and file structures are all about doing this. But, if you step back and look at the big picture, you quickly realize that such structured data constitutes 10% to 20% of the total data in any enterprise.

So, what about the remaining 80% to 90% of the data? What about the data and the information that's available within emails, paper or electronic documents (contracts, invoices, paystubs, purchase orders, loan origination forms, insurance claims forms, bank account initiation forms etc.) or the scanned images thereof, high speed camera feeds, voice files etc?

It's important to have the ability to not only, automatically and without human intervention, parse a document to generically classify it as a an invoice, a contract, a purchase order, a paystub, a plaintiff or defendant fact sheet, but to also be able to extract the relevant bits of information from these unstructured documents.

Those are the options made available by the technologies that convert unstructured data to structured formats. Once you have this data, your ability to automatically apply business rules enables a host of automation steps that cut significant processing time out of the process.

For example, this would apply to the human effort associated with:

- Determining compliance like regulatory, security or based on an organization's risk appetite
- Matching purchase orders and multiple invoices and bills of lading
- Matching data from the FICO score and investments statements and savings accounts and paystubs to a mortgage origination application and computing an applicant's qualification for a loan
- Determine the "sentiment analysis" associated with a document to determine the general media (social or otherwise) sentiment that is positive, neutral or negative about a person or organization

Q What are the most common challenges that you are seeing out there that people need to address?

A As it relates to unstructured data management the key challenges lie in the:

- Logical separation of the pages associates with a blob of data represented by a multi-page image or pdf
- Classification of these documents into their respective document types
- Machine learning based automated data extraction or capture
- Classification and data extraction from images and documents that are skewed or partial
- Recognizing document types that the machine learning models aren't familiar with and automatically adding them to the training models - thereby avoiding false positives and updating the machine learning training model
- Reducing the model training time through identification of similar document types for training and curation UIs which allow people to do their work in a business as usual model while automatically training the underlying model
- Handling variable length tables for data capture and extraction at commercial scale
- Applying a hybrid model that uses a combination of computer vision (visual features) and textual features to recognize document types and extract/capture data from them
- Machine learning models typically fail for use cases that require low volume/ high variability in document types with restrictions on retaining the training documents. Deep Learning neural networks need to be implemented to accommodate not just these use cases but also for creating enterprise level institutional memory that covers all typical and atypical use cases within a large enterprise

Q What is the difference between current / last gen tech as compared to next gen platforms out there?

A The difference between current and last gen tech is as follows:

Current / Last Gen Platforms:

- Could not automatically classify as they did not use computer vision
- Used traditional OCR and coordinate positioning as opposed to computer vision combined with OCR
- Did not have machine learning and AI integrated into the core of the system
- Could not use a hybrid of visual and textual features for classification and capture
- Created high error rates from “false positive” documents that the model wasn’t trained for
- Needed significant effort and a large sample set of documents to train the model
- Were tightly integrated with a specific OCR engine and hence their performance characteristics mirrored those of the OCR engine
- Could not handle data extraction from variable length tables
- Could not handle data extraction from skewed or truncated documents

Next Gen Platforms:

- Are able to use a combination of Computer Vision, Visual features and Textual features to classify documents automatically
- Use computer vision for classification and extraction
- Have machine learning and AI as integral core features of the platform
- Reduce the error rates from false positives by identifying the “novelty” documents and automatically routing them to the training environment
- Require a small sample set for training
- Are loosely coupled with OCR engines to be able to accommodate poor quality images with commercial grade OCR engines as and when required
- Are able to handle variable length tables for data extraction & capture
- Are able to extract/ capture data from skewed/zoomed/ truncated documents
- Some of the current gen platforms are also able to apply AI in the form of NLP/ deep or contextual learning and transfer learning models to enable evolutionary learning protocols that allow machine leaning and AI models to recognize similar but not the same (different) document types and automatically trigger the actions to handle the differences

IRPA AI | INSTITUTE FOR ROBOTIC PROCESS AUTOMATION & ARTIFICIAL INTELLIGENCE

For a limited time, the Institute for Robotic Process Automation & AI (IRPA AI) is putting together a 30-minute complimentary phone briefing on reducing costs by utilizing automation to streamline unstructured and semi-structured data.

To schedule your briefing, email Carrie.Simon@irpanetwork.com.

About Rohail Khan

Mr. Khan is a seasoned senior executive with over 25+ years of experience across multi-sector Fortune 1000 firms. His expertise lies within Banking, Insurance, Healthcare and BPO services transformation, automation, technology innovation, process reengineering, strategy development, international management and building global operating models. Mr. Khan is CEO of IRPA's Service Exchange, a firm that is engaged by Service providers, clients, Private Equity, Hedge Funds, Family Offices to assess companies across people, process and technology. Prior to that he was Founder and CEO of i-Tuple Inc. a technology services firm focused on global client optimization and transformation. Mr. Khan continues his work with investor groups that are evaluating and assessing BPO firms as well as clients who need assistance with developing BPO strategies and vendor selection.

About Frank Casale

Frank J. Casale is the founder and CEO of The Outsourcing Institute (OI) and the Institute for Robotic Process Automation and Artificial Intelligence (IRPA AI). Established in 1993, OI is a global marketplace and community of 70,000+ executive members including leading practitioners, service providers, advisors, thought leaders, industry observers and analysts. IRPA AI was established in 2013 as an independent professional association and knowledge forum for the buyers, sellers, influencers and analysts in the rapidly growing field of robotic process automation, cognitive computing and artificial intelligence.