

SPi Global: Content Extraction and Information Intelligence Suite

Businesses possess numerous sources of information that include documents, emails, databases, websites, and others. However, they fail to unlock business value out of the data without intelligent tools that can aggregate, normalize, semantically enrich, and convert data into actionable insights. To fulfill this need, SPi Global offers its Content Extraction and Intelligence Suite (CEIS) which harnesses machine learning (ML) and artificial intelligence (AI) technologies to decipher both structured and unstructured content—text, image, and voice—from different enterprise repositories and readies the data for analysis and search. “AI and ML help us to move away from a rules-based approach to automation/decisioning to a more cognitive approach. This unearths relationships and patterns from the content in multiple dimensions, helping clients discover new business opportunities and optimize processes,” remarks Ratan Datta, President and CEO, SPi Global.

“Most of the work we do involves technical and professional content where the combination of domain expertise and new age technologies is the key,” says Datta. For one of their engagements, highly qualified PhD SME experts identify and extract Chemistry and medical concepts reading research articles and patents and ingesting into the client’s database. The Concept Classification Tree has over 75,000 terms. “We have enhanced the manual task by incorporating an AI-based Recommender Engine to identify concepts automatically and suggest a classification to the operator. We see over 80 percent accuracy with the AI engine selecting terms,” adds Datta.

SPi sees an opportunity in helping clients manage regulatory compliance requirements across industries. For example, as a part of the MiFID regulations, financial institutions need to track and report on research information received. Using machine learning algorithms, SPi’s CEIS can read emails received and classify them into research, type of asset, and the category of firm. The content can be either in the email body, an embedded link or an attached document. The results are shown on a simple dashboard for compliance reporting, and management and businesses can plan their compliance strategy based on these results.

SPiCE 3G is a state-of-the-art document structuring and editing application within the CEIS framework. It can recognize patterns in deep SME content like scientific publications. SPiCE 3G uses pattern recognition technology



Ratan Datta
President & CEO

to identify key elements from the author’s manuscript for auto-structuring based on the Document Type Definition (DTD) schema of the client. Using natural language processing (NLP)-based techniques, the application has enabled a higher percentage of cross linking and semantic enrichment early in the publishing value chain. “We are now testing a language assessment module that can help determine the quality of language in a document,” says Datta. The module will help assign a document to a particular work queue for editing.

“We have been successful in using these technologies to improve operational excellence,” says Datta. In an instance, a publishing company sought a solution to reduce their time to market. SPi studied their workflow, identified potential areas of wait time, and created a new workflow that could bring the time down from four days to one. SPi also developed a module that could analyze the input material to predict a complexity and turnaround time score. With the module, the client could determine issues to be resolved before sending their content for production. Eventually, the module improved the one-day Turnaround Time (TAT) scores by over 20 percent.

SPi recognizes that handwritten text has always been a challenge and it looks forward to using AI technologies in improving content extraction for handwritten text. With its spate of success stories in mining key information from text and images, the company is now increasing focus on voice and video to help organizations realize the full potential of their content, making it more accessible, adaptable, and actionable. [CA](#)