Enabling Better Data Extraction from Tables in Contracts

cimplifi™

Do tables create clarity or complexity?

Legal contracts are full of tables. In a pre-digital era when contract drafting was paper-based, with human manual processing, lawyers used tables extensively to capture information concisely and display it effectively, providing information at any desired level of detail, precision, and hierarchy. Tables also made it possible to reduce the length of text, which is advantageous in lengthy legal contracts.

So, the presupposition is that tables exist to help humans make sense of complex data through simplified representations. **But are they that simple?**

Unfortunately, they aren't. How they are structured, and the data is represented can be vastly inconsistent, with no universal standard for creating a table for use in documents. Even intelligent people struggle to navigate and understand the information presented in tables. Those working in accessible document design believe that using tables for design purposes and reducing text length is generally considered bad design, creating both human accessibility barriers and barriers to adaptive technology.

Consequently, valuable contractual data is often hidden or obscured in tables in contracts and other documents that even humans struggle to consume and extract visually. This poses a significant risk when contractual or existential events dictate that critical information needs to be manually expedited for review from amongst a pool of other complex tabular data, particularly at volume.

Until recently, qualified analysts and paralegals have had to manually locate and extract tabular data, which is slow, inefficient, risky, and costly. With the layout of tables formatted for visual consumption by humans, not technology, barriers exist, which means that computers couldn't do the task accurately. **Let's understand why and how it has forced technology to evolve.**

The technical challenge

A table is much more than its data.
It is already difficult to extract data with technology; tabular data compounds the challenge threefold. Why?

1. Varying table structures

    - Tables within the same document type can have varying structures. Data can be semi-structured and unstructured within a table, with inconsistent data point locations.
    - Information is often not uniform throughout the table, and cell content uses different syntactic representations such as symbols, text, abbreviations/acronyms, numbers, and mathematical notations. Tabular extraction, therefore, requires knowledge of all possible presentation patterns.

2. Context and hierarchy

    - Contextual information and hierarchical structure are often lost in an extracted table and table data, which has historically meant that analysts needed to restructure extracted table data and add context manually.
    - Unstructured text segments, such as table titles, footnotes and non-table prose discussing a table, are often required to interpret the table's entries.
    - Headings in documents are used as structural elements to provide information on the logical order of content and to navigate to a specific topic or sub-topic in a document, not a particular cell.
    - Including headings in a table forces humans and technology to navigate to a particular cell that does not correlate to a topic or sub-topic in a document.

3. Format

    - The document format containing the table can vary e.g. .pdf, text, HTML etc. Some formats are more challenging than others, e.g. .pdf has no internal representation of a table structure.
    - Tables can use different languages or industry-specific jargon.
    - Information and data contained within a table are displayed multi-dimensionally using a two-dimensional linear format. A set of data and data context is presented in a non-standard format.

The evolution of table data extraction technology

Fortunately, technology-driven solutions now exist to circumvent these human barriers, thanks to optical character recognition (OCR), artificial intelligence (AI) and machine learning.

Table data extraction uses a combination of OCR, AI and machine learning models in a self-learning system that automates the detection, recognition, and extraction of table data from contracts.

By training datasets on representative sample contracts, AI models are built and designed to isolate and extract specific table data to precondition AI-powered searches. A tabular data model normalizes, calibrates, classifies, and validates data for consumption in downstream systems or form creation in a contract lifecycle management (CLM) solution.

Humans can then use exception management processes to identify partial or non-automatable table datasets, reviewing each in detail to develop case studies. Machine learning modules are trained in a self-learning cycle by leveraging these information-rich case studies.

The future

In the future, intelligent contracts will eradicate the challenge of tables of valuable contract data being locked in physical contracts, thereby removing the human barrier from the equation entirely.

A foundation of industry-led data standards in digital formats will lead to structured, transparent, and consumable intelligent contracts that provide uniformity in machine-readable text.

Developing intelligent contracts with adaptive technology and AI-based analytics and solutions in mind will further accelerate existing solutions.

To learn more about the CALM™ practice at Cimplifi and our expertise in table data extraction, contract analytics, and lifecycle management click here.

#keepCALMandcontracton

cimplifi.com | calm@cimplifi.com