



# Data Extraction for Semi-Structured Documents

**EXTRACT VALUABLE DATA AND INSIGHTS FROM PDFS, WORD DOCUMENTS, AND MORE**

*Lexalytics is the industry leader in translating text into profitable decisions through the Lexalytics Intelligence Platform, a complete, modular data analytics solution.*

*Data analysts turn to Lexalytics to gain more value from large volumes of structured, unstructured and semi-structured text documents.*

## Structured data

Structured data lives in fixed locations inside of spreadsheets or databases. This setup gives the data an “address” where it can be found, such as row or column. Business intelligence (BI) tools use this location in order to identify, reach and extract relevant data for each query. Structured data is what every data analyst dreams of: neat, organized, and ready to be analyzed.

## Unstructured data

Unstructured documents, such as tweets and news articles, also contain valuable data. But data within these documents don’t have “addresses” in the same way that a spreadsheet cell has a location. Individual data points, like companies or product names, are free-floating within the raw text. Natural language processing (NLP) systems analyze this unstructured text documents to identify and organize the data within.

## SEMI-STRUCTURED DOCUMENTS:

- Contracts
- Regulatory updates
- Research papers
- Financial documents
- Market reports
- News articles
- SEC filings
- Requests for Proposal (RFPs)

## Semi-structured data

Some documents contain structured data hidden in a seemingly unstructured

format. Think about legal contracts, financial documents, or even this datasheet. These are written in text but have structured sections organized under headings or contained in tables, bullets and lists. Each section adds a layer of context to the words within. For example, it’s useful to know that a contract mentions a particular company and a dollar amount. But it’s even more valuable to know that these are mentioned in the context of “service level agreement.”

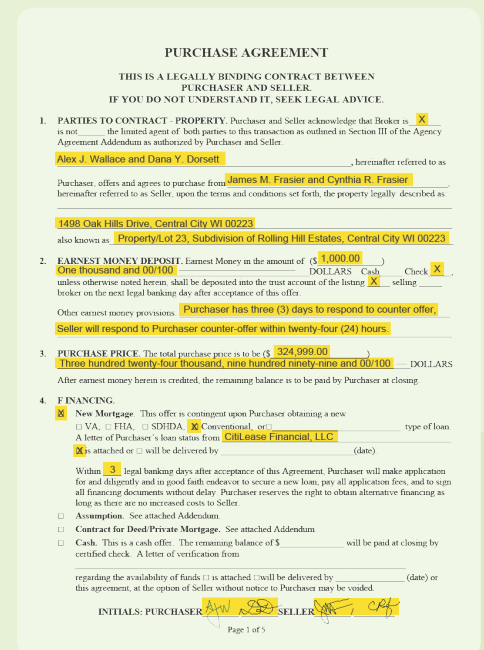


Figure 1 | Example of a semi-structured document: real estate purchase agreement

**CONTACT US TODAY**



## Applications of data extraction services:

- Help customer support agents resolve issues more quickly
- Analyze data from stock market reports to identify trending companies
- Stay up-to-date with regulatory updates and changes
- Flag input errors and suspicious financial recommendations for an auditor to review
- Extract relevant information from EHRs to improve clinical decision making and revenue cycle management
- Aggregate similar contracts to comply with Public Contracts Regulations

Extract  
**VALUABLE DATA**  
and  
**GAIN CONTEXTUAL INSIGHTS**  
from semi-structured documents

BI tools are built to handle structured data but stumble with unstructured text because each sentence looks like one very large datum. NLP tools excel at analyzing unstructured text, but don't consider the importance of structure. This leaves semi-structured documents in a "dead zone" where BI and NLP tools both fall short.

### VALUABLE DATA TO EXTRACT:

- Recommendations
- Key Opinion Leaders
- Customer requirements
- Subscription details
- Disclaimers
- Deadlines
- Stock ticker symbols
- Age ranges
- Codes
- Products
- Cash amounts
- Order numbers
- Illnesses
- Companies and brands

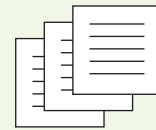
### Unlock the full value of your documents

Lexalytics solves this problem by using our deep experience with text data in all its forms to evaluate the underlying structure of your semi-structured documents. Once we understand the pieces, we extract structured data directly while using our natural language processing software to transform unstructured text into structured data. Where traditional BI tools would miss this data entirely, our approach delivers a new layer of contextual insights to you and your customers and opens up a range of broader analytics applications.

**CONTACT US TODAY**

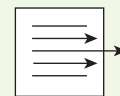


## 4-STEP DATA EXTRACTION SERVICES



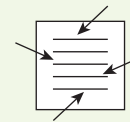
### 1 Gather documents

You have valuable data locked away in complex, semi-structured text documents



### 2 Evaluate structure

We use our understanding of text in all its forms to evaluate the underlying structure of your semi-structured documents



### 3a Extract structured data

We automatically extract already-structured data and insert it into your database, warehouse or other storage system



### 3b Process unstructured data

We use our natural language processing software to transform unstructured text into structured data, adding a new layer of contextual insights



### 4 Integrate or act

Add this analytics capability into your products for customers to leverage, or improve your decision-making

Figure 2 | Lexalytics 4-Step Data Extraction Services helps data analysts transform previously-inaccessible information into structured data and useful insights