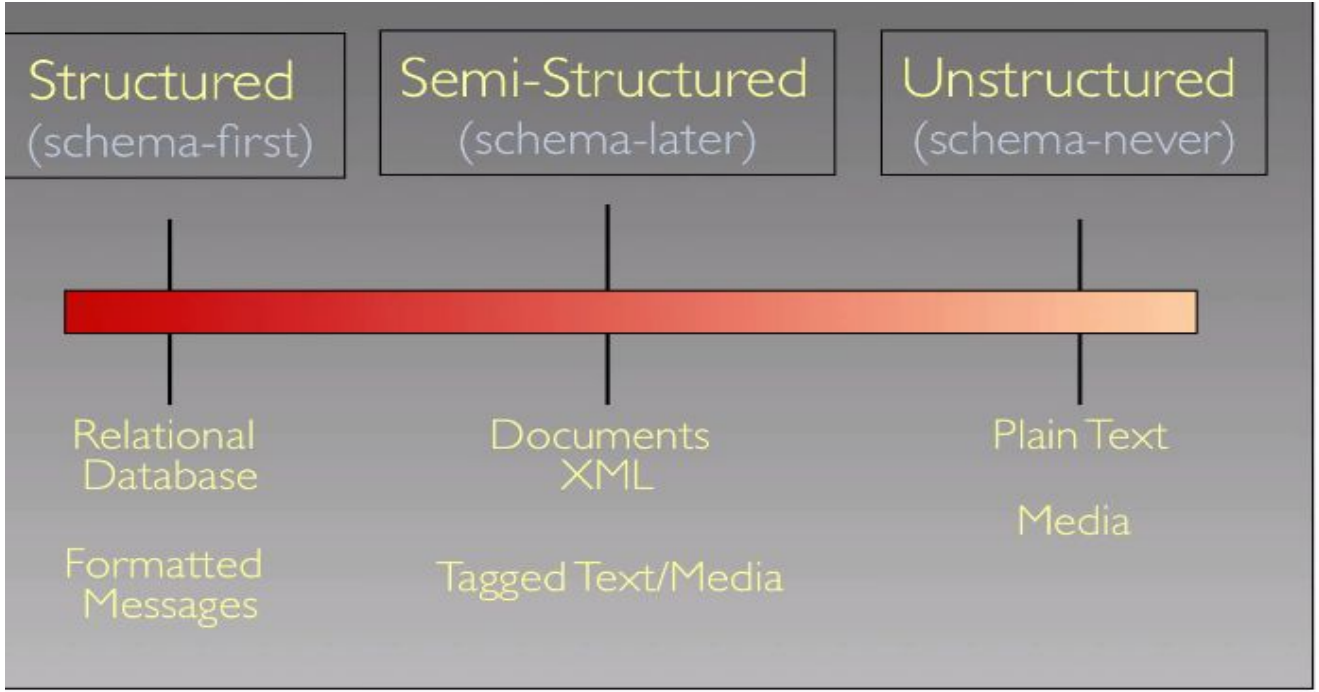
Text Extraction from PDFs made Easy!

Aakriti Jain • Automation Specialist: S&P Global: Market Intelligence
aakritijain96@gmail.com • https://github.com/Aakriti23

Types of Data



Structured Data

- Data stored in tables with rows and columns. This data contains well-defined data types and column headers that make it easier for machines as well as humans to read.
- Common formats to store structured data - RDMS databases like SQL Database.

OID	XPATH	DATA	MODEL
1	order/customer/address	9	1
2	order	9	1
3	order/qty	9	1
4	order/customer/address/street	9	1
5	date	9	1
6	status	9	1
7	order/ordernr	9	1
8		9	1
9	order/customer/name	9	1
10	order/customer	9	1
11	customer/address	10	1
12	qty	10	1
13	ordernr	10	1
14	customer/name	10	1
15	customer	10	1
16	customer/address/street	10	1
17		10	1

Semi-structured Data

- Information doesn't reside in a relational database but it does have some organizational properties that make it easier to analyze.
- Common formats to store semi-structured data - XML, JSON, CSV etc.

XML Example

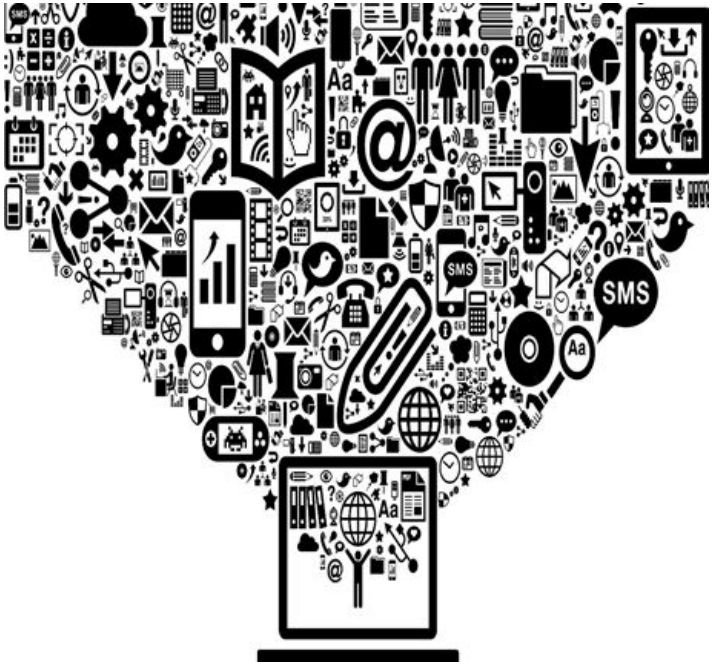
```
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
        <food>
        <name>Belgian Waffles</name>
        <price>$5.95</price>
        <description>Two of our famous Belgian Waffles with plenty of real maple syrup</description>

        <calories>650</calories>
        </food>
        <food>
        <name>Strawberry Belgian Waffles</name>
        <price>$7.95</price>
        <description>Light Belgian waffles covered with strawberries and whipped cream</description>

        <calories>900</calories>
        </food>
        <food>
        <name>Berry-Berry Belgian Waffles</name>
        <price>$8.95</price>
        <description>Light Belgian waffles covered with an assortment of fresh berries and
whipped cream</description>
        <calories>900</calories>
        </food>
</breakfast_menu>
```

Unstructured Data

- This might/might not have internal structure but is not structured via pre-defined data models or schema. It may be textual or non-textual, and human- or machine-generated.
- Typical unstructured data includes - Text files (Word documents, PDFs, .txt files), Email body, Social Media (Facebook, Twitter, etc.)



Problem Statement

We receive numerous bank letters pertaining to information about ADR/GDR (American Depository Receipts & Global Depository Receipts) on a shared mailbox.

The number of emails can be averaged out to 200 emails/week.

Researcher manually goes to each email and downloads the PDF attachments.

She/he then copies roughly 18 data fields from each PDF into an excel table which later gets inserted into the tool.

Cons of doing the work manually

- Too many man-hours consumed.
- Copying sensitive information manually increases chances of error.

We can utilize one Python library to make the extraction from documents easy and reduce manual efforts.

Let’s begin with Regular Expressions

Introduction to Regular Expressions

- Regex (also called Regular Expression) is a mini-language that looks encrypted and mysterious at first.
- It is a wildcard that helps in parsing through strings and matching exact patterns in a text.

- Regex can assist in a wide variety of tasks such as:
  - Text matching
  - Repetition
  - Pattern-composition