DATA EXTRACTION

In today's data-driven world, the ability to extract, process, and utilize information from multiple sources efficiently is crucial. Businesses, researchers, and professionals frequently work with data in varied formats such as images, scanned documents, PDFs, Excel sheets, and more. The process of extracting data from these sources can be streamlined using various tools and technologies, including Optical Character Recognition (OCR), machine learning, and automation frameworks.

1. Data Extraction from Images

Extracting data from images involves recognizing and converting textual content into a structured format. This is particularly useful when dealing with screenshots, scanned documents, or or infographics.

Methods:

i. Optical Character Recognition (OCR): Tools like Tesseract OCR, Google Vision API, and Adobe Acrobat can recognize text in images and convert it into editable formats.

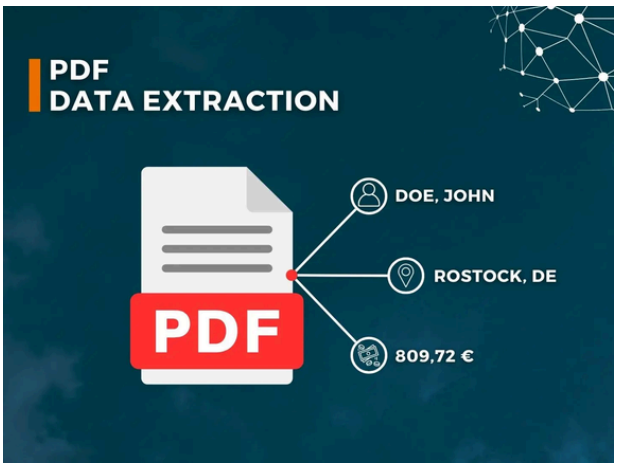
ii. AI-powered Image Processing: AI models can extract tabular data, handwritten text, and even numeric values from charts or graphs.

iii. Pre-processing Techniques: Enhancing contrast, noise reduction, and edge detection improve accuracy in OCR-based extraction.

Applications:

i. Extracting invoices, receipts, and legal documents.
ii. Digitizing handwritten notes or old manuscripts.
iii. Converting infographics into structured datasets.

2 . Data Extraction from PDFs



Portable Document Format (PDF) is widely used for reports, contracts, and forms. Extracting data from PDFs requires different techniques depending on whether the document is text-based or image-based.

Methods:

i. Text-Based PDFs: Tools like PyPDF2, PDFMiner, and Tabula can extract text and tables from machine-readable PDFs.

ii. Scanned PDFs (Image-based PDFs): Require OCR-based processing using Tesseract or ABBYY FineReader.

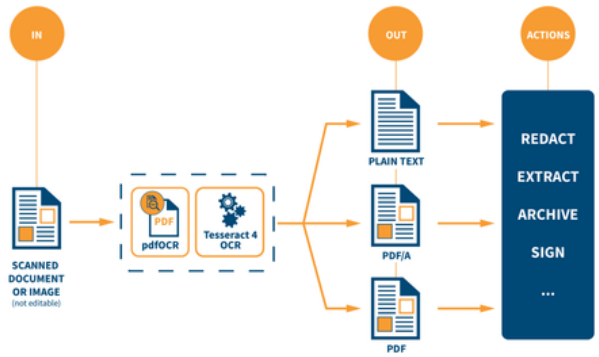
iii. Automated Extraction: Python libraries like Camelot and PDFPlumber extract structured tabular data with high accuracy

Applications:

i. Extracting financial reports, legal documents, and academic research papers.
ii. Automating data collection from business contracts.
iii. Parsing structured data from invoices and forms

3. OCR-Based Data Extraction from Scanned PDFs

Scanned PDFs, being image-based, require specialized OCR technology for text recognition.



Steps in OCR Processing:

i. Pre-processing the Document: Enhancing the scan quality, removing distortions, and applying thresholding.
ii. Text Recognition: Using OCR tools like Tesseract, Google Vision, or Amazon Textract.

iii. Post-processing: Cleaning up errors, structuring text, and exporting to usable formats (Excel, CSV, JSON)

Challenges:

i. Low-resolution scans can reduce OCR accuracy.
ii. Complex layouts, handwritten text, or faded prints require additional AI-based enhancement.

4. Data Extraction from Excel Files

Excel remains a widely used format for storing and managing structured data. Extracting data from Excel files can be done programmatically for automation and data analysis.

Methods:

i. Python Libraries: Pandas, OpenPyXL, and xlrd help extract, modify, and process Excel data.

ii. Power Query (Excel Tool): Used for cleaning and transforming raw data.

iii. VBA Automation: Macros can automate extraction and formatting tasks.

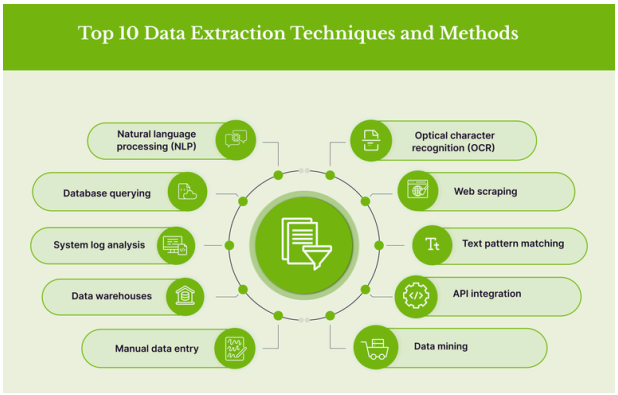
Applications:

i. Extracting financial statements and reports.

ii. Automating reconciliation tasks in accounting.

iii. Consolidating data from multiple Excel files.

5. Extracting Data from Other Input Sources



Apart from the major formats discussed, data extraction is also required from emails, web pages, and APIs.

Emails:

i. Email parsing tools (MailParser, Zapier) extract structured data from emails.

ii. Python's imaplib and email libraries help automate email processing.

Web Scraping:

i. Tools like BeautifulSoup and Scrapy extract data from web pages.

ii. Selenium is useful for extracting dynamic content from websites.

APIs & Databases:

i. APIs allow seamless data retrieval using JSON or XML formats.

ii. SQL queries help extract structured data from relational databases.

6 PRACTICAL USE CASES

6.1 Instant Data Extraction from Image to Table Format

I had an image containing data in a tabular format with around **60 rows.** If I had manually typed it, it would have taken me at least **one hour.**

Instead, I uploaded the image to **ChatGPT** and simply asked it to extract the data **exactly as it appeared** in the image.

You won't believe it—**within a minute**, I got the entire data perfectly formatted in a table!

I copied it and used it for my purpose effortlessly. Even when I pasted it into **Excel,** the data automatically fit into separate columns without any issues.

This saved me so **much time and effort!**

6.2 Extracting Bank Statement Data from PDF with Proper Formatting

I had a **10-page bank statement PDF**, where the **debit (Dr) and credit (Cr) amounts were scattered** across different lines without a clear tabular structure.

Manually copying and organizing this data would have been a **time-consuming** and **error-prone** task.

Instead, I uploaded the **PDF** to **ChatGPT** and asked it to:

i. Extract the data **exactly as it appeared** in the statement.

ii. Present it in a **clean tabular format**.

iii. **Separate the Dr and Cr amounts into different columns** for better clarity

Within **seconds,** I got a well-structured table with all transactions properly formatted! I was able to **copy and use it effortlessly,** even in **Excel,** where the Dr and Cr amounts were neatly placed in separate columns.

This saved me **hours of manual work** and ensured **100% accuracy** in the data!

6.3 Extracting Sales Data from Excel as per Specific Requirements

I had a large **Excel file** containing extensive **sales data,** including **transaction dates, product names, quantities sold, unit prices, and total sales amounts.** However, I needed to extract only **specific information** based on my analysis requirements.

For example, I wanted to:

i. Extract sales data **only for Smartphones and Laptops.**

ii. Filter transactions **within the last three months.**

iii. Show **only those sales where the total amount exceeded ₹50,000.**

If I had manually applied filters and organized the data, it would have taken me **a lot of time and effort.**

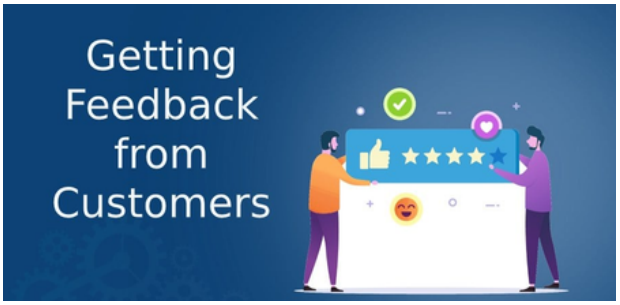
i. Extract only the **relevant columns and rows** I needed.

ii. Filter data based on **my specified conditions.**

iii. Present the extracted data in a **clean and structured format.**

Within **seconds,** I got the exact sales data I was looking for—**well-organized and ready to use.** I even copied it into **Excel,** and it fit perfectly into separate columns, **saving me hours of work!**

6.4. Extracting Customer Feedback from Surveys



A company conducted an online survey with thousands of customer responses stored in a **CSV-Excel file.** The responses included both **structured (rating□based) and unstructured (open-ended comments) data.**

Problem:

Manually reviewing and categorizing feedback based on sentiment, common issues, or product preferences would take **days.**

Solution:

With **ChatGPT,** the company:

i. Extracted only **negative reviews** for product improvement.

ii. Identified **common keywords** and **recurring issues** automatically.

iii. Categorized feedback into **themes like pricing, quality, and customer service.**

Result: Within **minutes**, they had an actionable report on customer concerns!

6.5. Extracting Transaction Details from Email Statements

A freelancer receives **monthly PayPal transaction emails** and needs to log income details into an **Excel sheet** for tax filing

Problem:

Manually copying **transaction date, sender name, amount, and currency** from multiple emails is **time-consuming and prone to errors.**

Solution:

With **ChatGPT,** the freelancer:

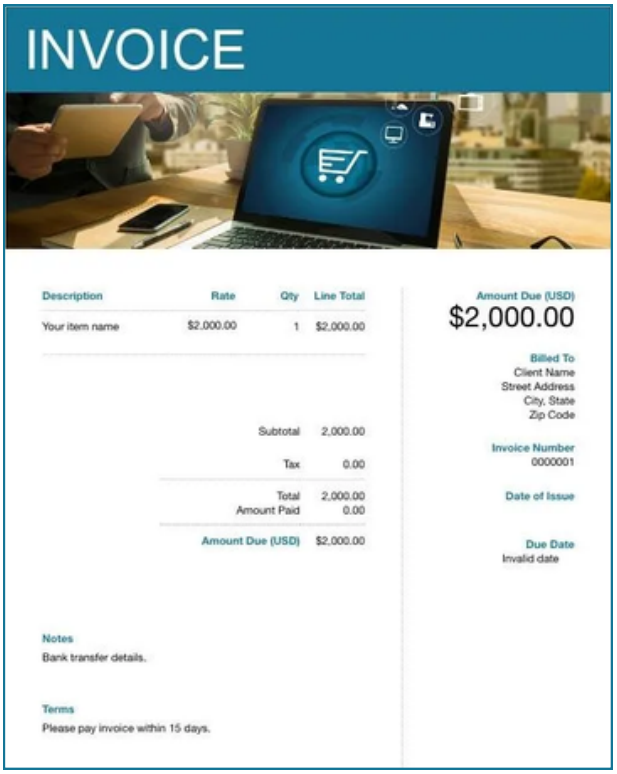
i. Uploaded a set of **email text extracts.**

ii. Asked ChatGPT to **extract and format key details**.

iii. Copied the structured data directly into Excel.

Result: Hours of manual work was **done in minutes** with no mistakes

6.6. Extracting Order Details from E-Commerce Invoices (PDFs)



A business owner frequently receives **supplier invoices in PDF format** but needs to extract **order details** (product names, quantities, and prices) into an **inventory management system.**

Problem:

Invoices have **different formats,** and manually entering data would be **time consuming.**

Solution:

By using **ChatGPT,** the owner:

i. Uploaded **multiple PDFs** for extraction.

ii. Received a **clean table with order details.**

iii. Copied the data directly into their **inventory tracking software**

Result: Inventory updates became **instant and hassle-free!**

6.7. Extracting Key Data from Legal Documents

A law firm needs to review **contract agreements** to extract specific terms, dates, and clauses.

Problem:

Contracts are lengthy, and manually locating key clauses (e.g., **termination policy, renewal terms, and payment obligations)** is **tedious.**

Solution:

With **ChatGPT,** they:

i. Uploaded **PDFs of contracts.**

ii. Asked ChatGPT to **identify and extract important clauses.**

iii. Received a structured summary for quick reference.

Result: Legal reviews were **sped up significantly,** reducing research time!

.6.8. Extracting Product Listings from a Website

A digital marketer needs to collect product details (names, prices, ratings) from an **e-commerce website** for competitor analysis.

Problem:

Manually copying product information **from hundreds of web pages** would be **extremely slow.**

Solution:

With **ChatGPT** and web scraping tools, they:

i. Extracted **product names, prices, and reviews** automatically.

ii. Filtered data by **category, brand, and discount range.**

iii. Received a structured **Excel-ready** format.

Result: A **comprehensive competitor analysis** was ready in minutes!

6.9. Extracting Financial Metrics from Annual Reports

An investment analyst needs to extract **key financial metrics (revenue, profit, assets, liabilities)** from **company annual reports (PDFs)** for analysis

Problem:

Finding and copying these numbers from **hundreds of pages** is **time□intensive.**

Solution:

With **ChatGPT,** they:

i. Uploaded **PDF reports** and asked for financial summaries.

ii. Extracted **profit & loss, balance sheet, and cash flow** statement data.

iii. Copied the structured data into **Excel for further analysis.**

Result: Faster **financial statement analysis** for better decision-making!

6.10. Extracting Meeting Minutes from Recorded Transcripts

A project manager needs to extract **key action points** from **meeting transcripts (text format from Zoom-Teams recordings).**

Problem:

Reviewing long transcripts and summarizing action items manually is **time-consuming.**

Solution:

With **ChatGPT,** they:

i. Uploaded transcripts and asked for a **summary of key points.**

ii. Extracted **decisions, tasks assigned, and deadlines.**

iii. Got a **well-organized meeting summary** in seconds.

Result: Instant meeting notes without manual effort!

6.11. Extracting Expenses from Scanned Receipts

A small business owner wants to digitize **paper receipts** for tracking expenses.

Problem:

Manually entering **dates, vendors, and amounts** from hundreds of receipts is **slow and inefficient.**

Solution:

Using **ChatGPT with OCR,** they

i. Scanned receipts and extracted **key expense details.**

ii. Got a **structured Excel file** with all expenses categorized.

iii. Used the data for **expense tracking & tax reporting**

Result: Automated bookkeeping with no manual data entry!

CA inderjeet Kaur Bamrah

Inderjeet Kaur Bamrah is a visionary Chartered Accountant, distinguished author, and a passionate advocate for the convergence of finance and artificial intelligence. With a deep understanding of financial reporting, corporate compliance, and business process automation, she is committed to empowering professionals with the knowledge to navigate the rapidly evolving technological landscape.