AutismConnect

Data acquisition, extraction, storage Project

Pr. SENELLART Pierre

2025 /2026

Dauphine | PSL
UNIVERSITÉ PARIS

AUTISM CONNECT

Team members

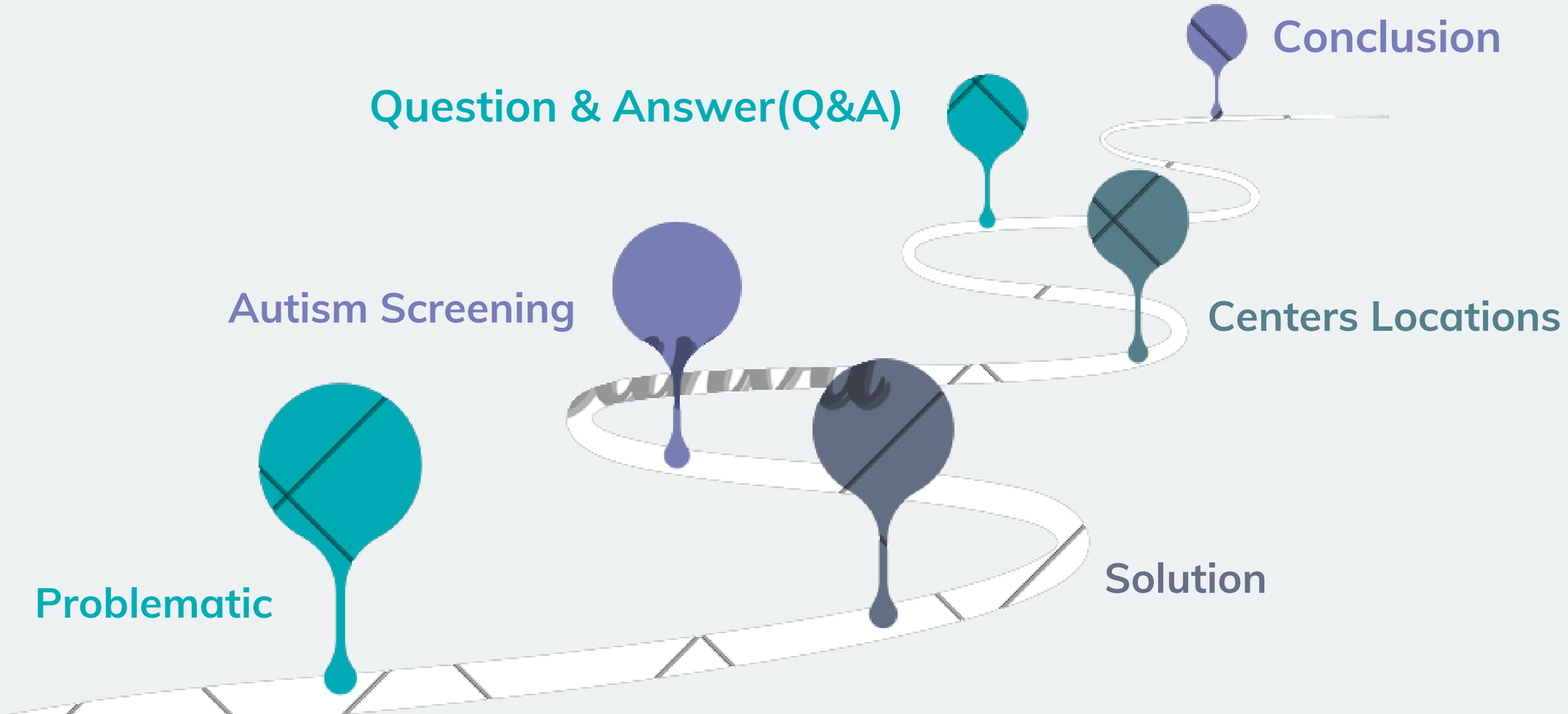
LAFIFI Bochra

BOUMAZOUZA Ines Manel

CHOUCHOU Fatma Ibtissam

AUTISM CONNECT

Table of contents



The struggle is real

- 1 in 36 children is diagnosed with autism spectrum disorder (ASD).
- 1 in 45 adults in the U.S. live with this condition.
- Challenging conditions :
  - (8% of autistic students in the U.S. don't finish high school.
  - Only 21% of people with disabilities, including autism, are employed.

Unveiling the Solution:

Solution : The Power of AutismConnect

Autism Detector:

Analyse specific information provided by parents to assess the likelihood of autism in their child.

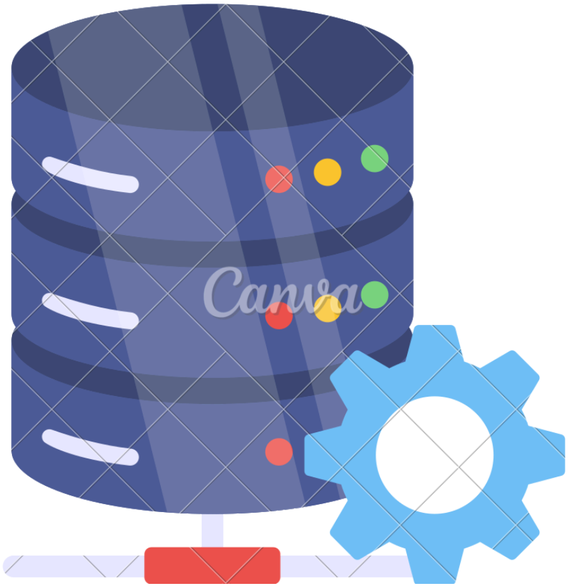
Autism Center Locator:

Provides Parents with Nearby Autism Centers Based on Their Location.

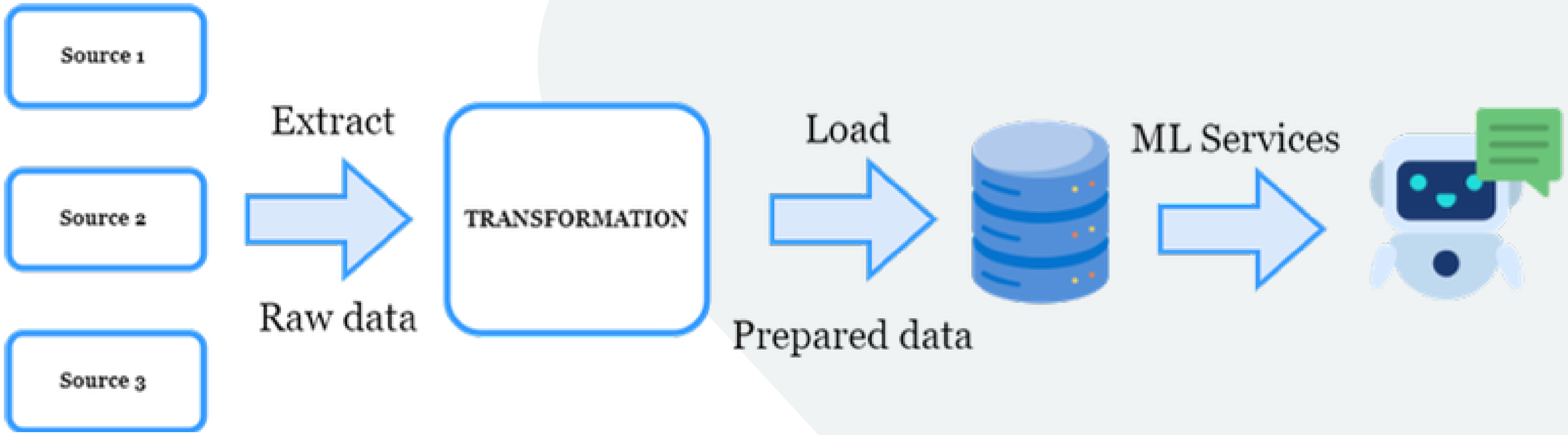
Autism companion:

Provides tailored guidance and individualized recommendations to effectively tackle the challenges linked with autism.

Behind the Scenes: the Data Infrastructure



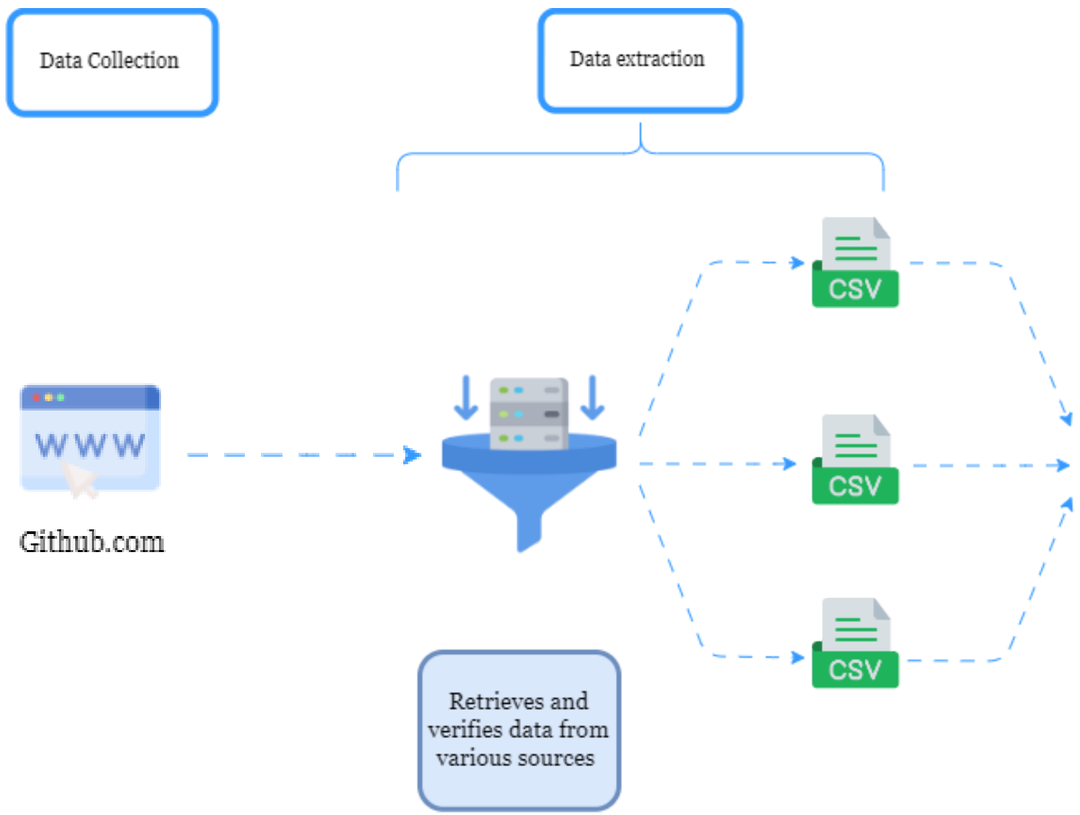
General Pipeline



AUTISM CONNECT

Autism Screening

Autism Screening

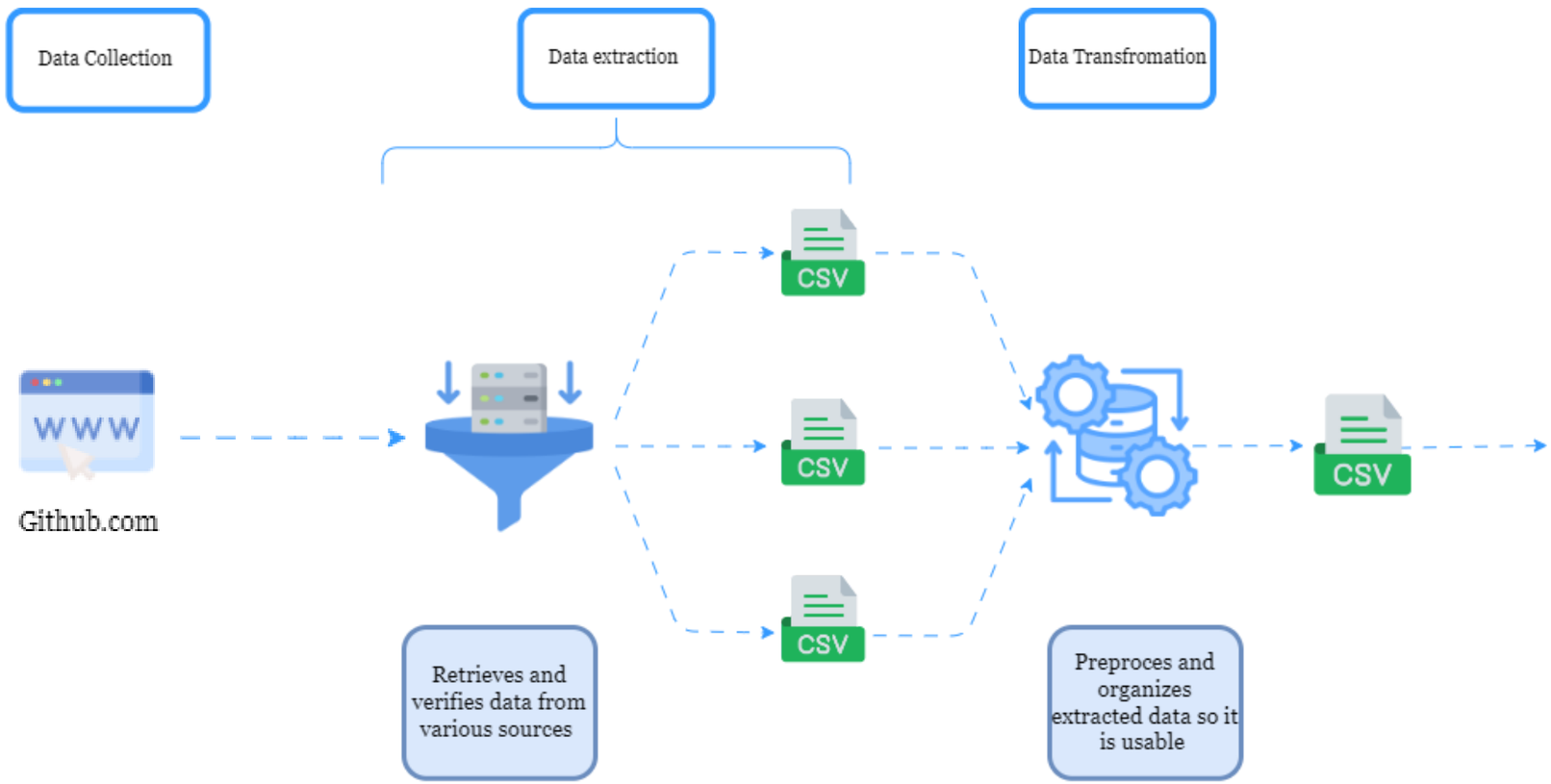


GitHub Repository - Autism Diagnostics Dataset

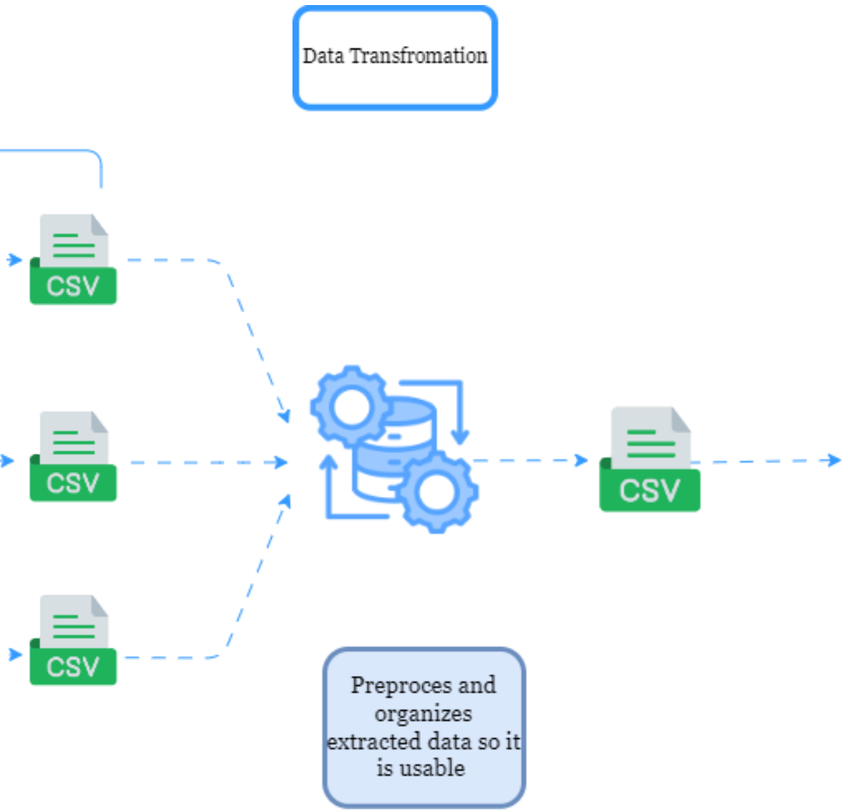
- **Source**: shaheennamboori/CSV_dataset_for_autism_diagnostics
- **Contains 3 CSV files**: Child, Adolescent, Adult screening data
- Based on AQ-10 (Autism Spectrum Quotient) diagnostic test
- 1,100 total records from 82 countries
- Direct CSV download via GitHub raw URLs
- **Decision:** Use Python requests library for automated HTTP download

https://github.com/shaheennamboori/CSV_dataset_for_autism_diagnostics

Autism Screening



Autism Screening



- **Column standardization:** Lowercase, fix typos (contry_of_res → country)
- **Duplicates removal**: 9 rows eliminated across datasets
- **Missing value handling**: Remove 142 incomplete records (marked as ?)
- **Country name standardization:**
  Fix variations (americansamoa → american samoa)
- **Data type conversion:**
  - Screening scores (A1-A10) to integers (0/1)
  - class_asd: 'yes'/'no' → binary (0/1)
- **Feature engineering:** Add age_group column (child/adolescent/adult)
- **Dataset merging:** Combine 3 CSV files into unified dataset

Output: **946 clean rows × 20 column**

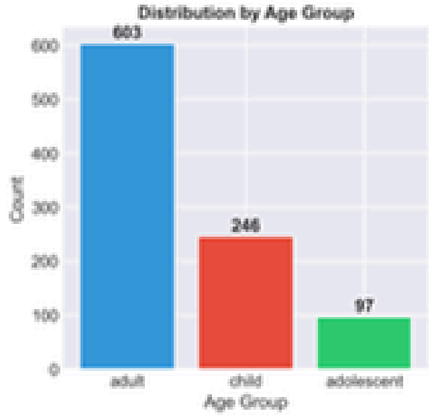
Autism Screening

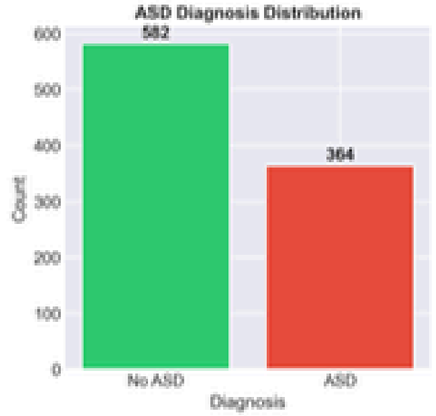
Data Quality Assessment

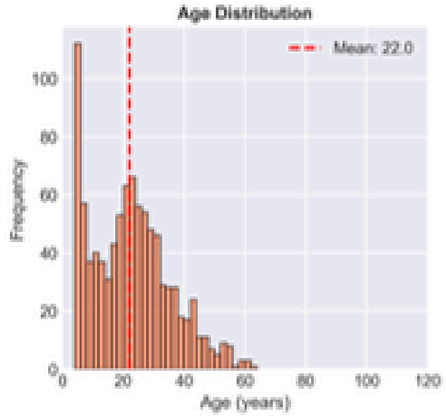
Quality Check Methodology

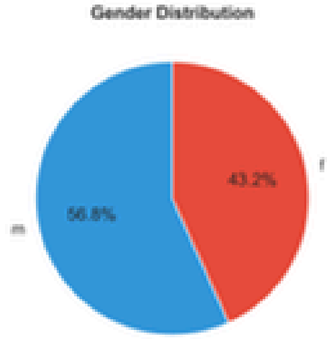
- Missing values: Handled by row removal
- Duplicates: 0 (all removed)
- Outliers: Age > 150 removed (1 row)
- Invalid data: 0 (all validated)

Metric	Before	After
Total rows	1,1	946 ✓
Missing values	242	0 ✓
Duplicates	9	0 ✓
Invalid ages	1	0 ✓

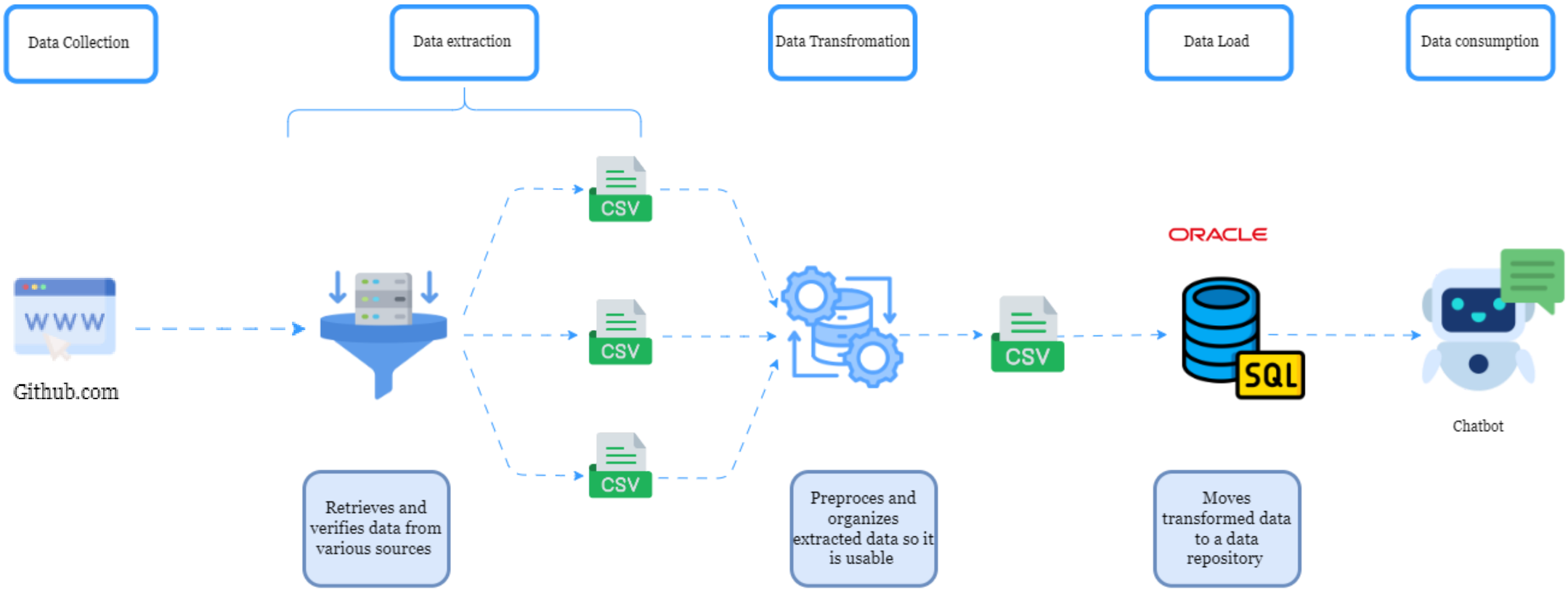








Autism Screening



Autism Screening

Data Storage

- **Future data integration**: Link screening data with autism centers location via country field for geographic analysis
- **Data integrity for medical data:** Enforce constraints (age limits, valid categories) and prevent duplicates through database triggers
- **Structured data with relationships:** Fixed schema ensures consistency and supports complex queries across datasets

ORACLE

Autism Screening

Database Testing & Quality Checks

1. Auto-increment ID Trigger

```
INSERT INTO autism_screening (A1_Score,...., age_group, Class_ASD)
  VALUES (v_A1_Score,...., v_age_group, v_Class_ASD)
  RETURNING id INTO v_test_id;
  DBMS_OUTPUT.PUT_LINE('Insertion réussie. ID de test: ' || v_test_id);
```

Result:
Insertion réussie. ID de test: 947

2. Validation Queries

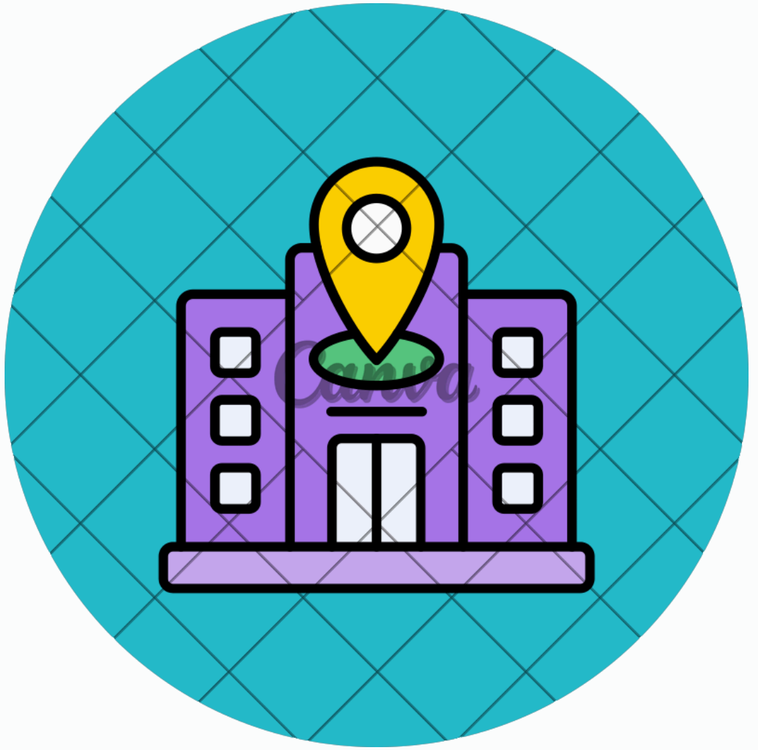
```
  -- Verify table structure
SELECT column_name, data_type FROM user_tab_columns
WHERE table_name = 'AUTISM_SCREENING';
-- Count loaded records
SELECT COUNT(*) FROM autism_screening;
```

Result:
20 columns verified | 946 records loaded

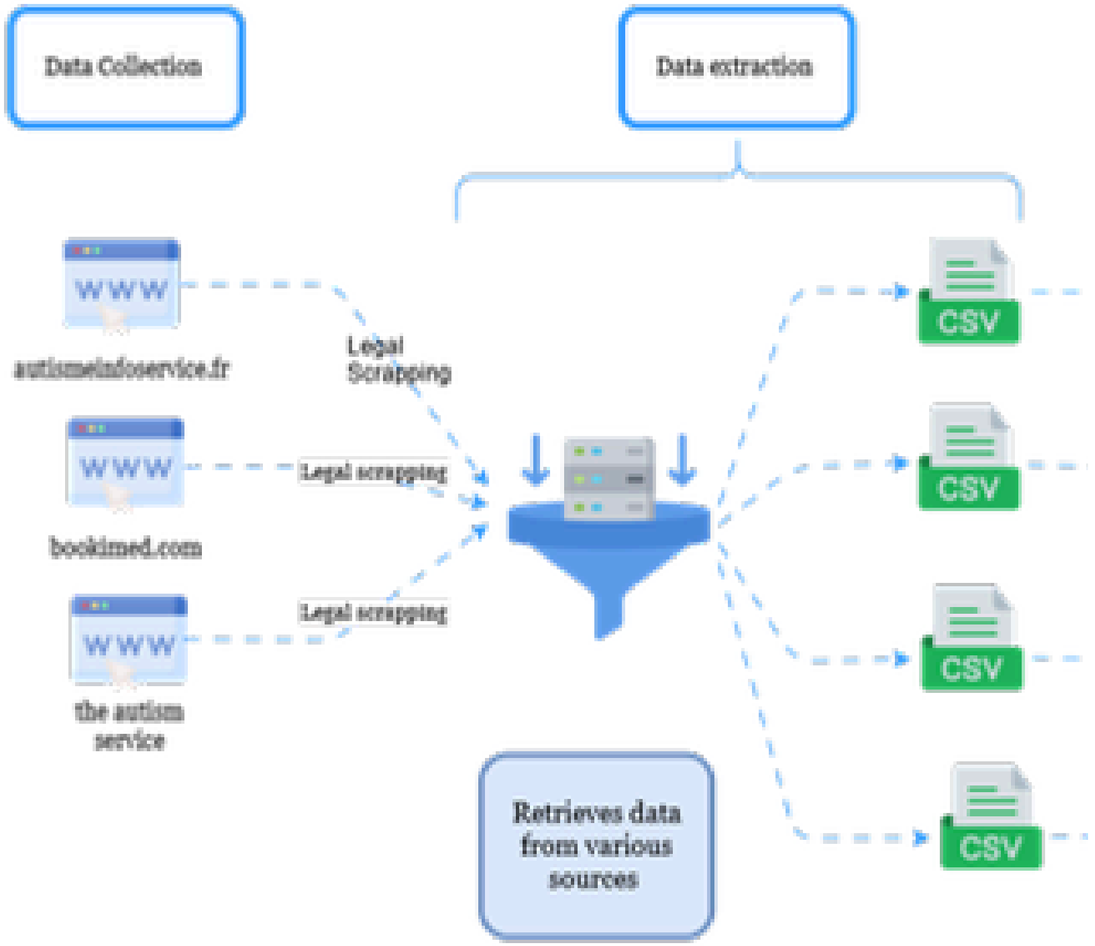
AUTISM CONNECT

Autism Centers Locations

EA



Centers Locations



Source 1: Autisme Info Service (French Government Directory)

- Provides structured and paginated listings of autism diagnostic and resource centers across France.
- Python-based scraping using requests library for HTML **retrieval** and **BeautifulSoup** for parsing
- Script iterates through all result pages by dynamically modifying the page index in the URL
- **Delays** introduced between successive requests to respect server load
- Extraction continued until no additional results detected, ensuring complete data collection

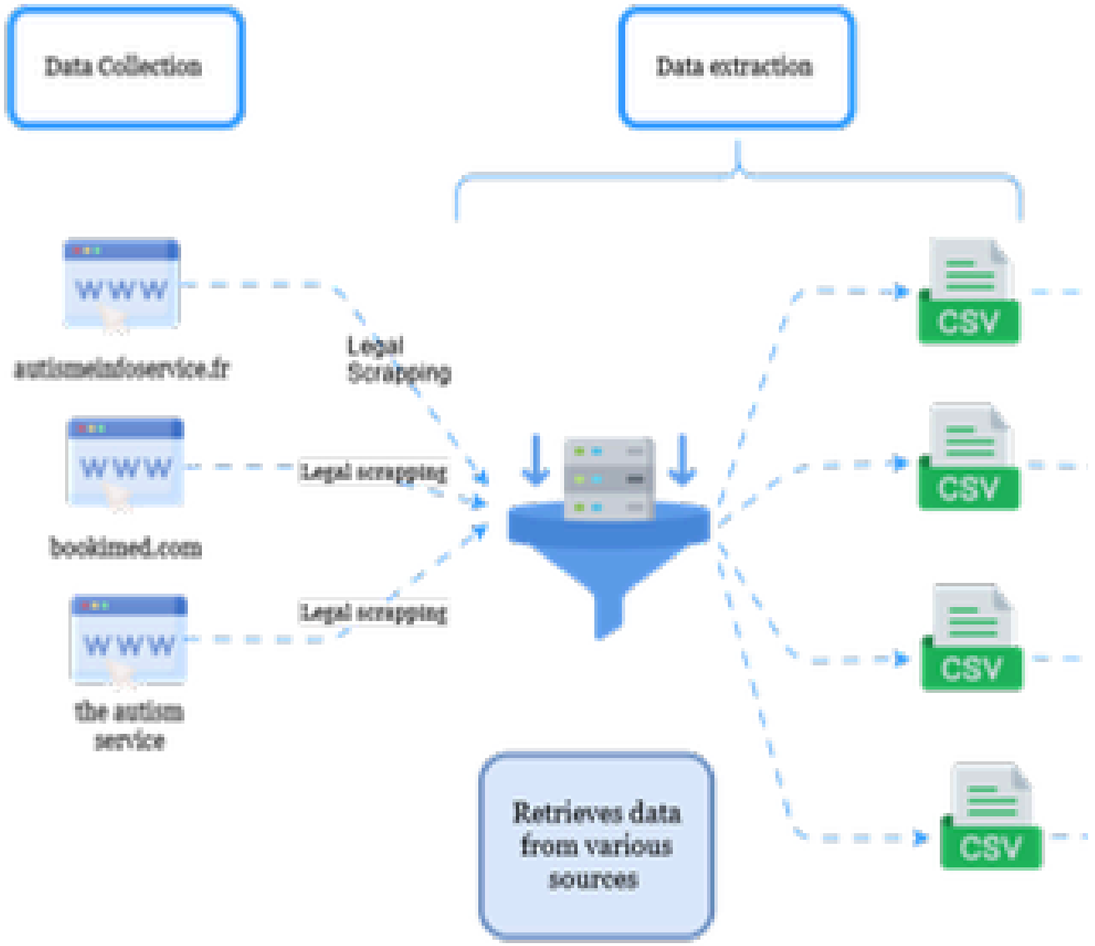
Centers Locations

Source 1: Autisme Info Service (French Government Directory)

```python
base_url = "https://annuaire.autismeinfoservice.fr/recherche/rubrique/rubrique:14/
#rubrique:14 for Diagnostic center, 10 for ressources cente
data = []
page = 1

while True:
    print(f"Scraping page {page}...")
    response = requests.get(base_url.format(page))
    soup = BeautifulSoup(response.text, "html.parser")

    # Find all table rows
    rows = soup.find_all("tr")

    # Filter rows that actually contain center info
    center_rows = [row for row in rows if row.find("td", class_="padding_left")]
```

```python
for row in center_rows:
    cols = row.find_all("td")
    center_name = cols[0].get_text(strip=True)
    region = cols[1].get_text(strip=True)
    address = cols[2].get_text(strip=True)
    phone = cols[3].get_text(strip=True)
    details_link = cols[4].find("a")["href"] if cols[4].find("a") else ""
    data.append({
        "Center Name": center_name,
        "Region": region,
        "Address": address,
        "Phone": phone,
        "Details Link": "https://annuaire.autismeinfoservice.fr" + details_link
    })

page += 1
time.sleep(1)  # polite delay
```

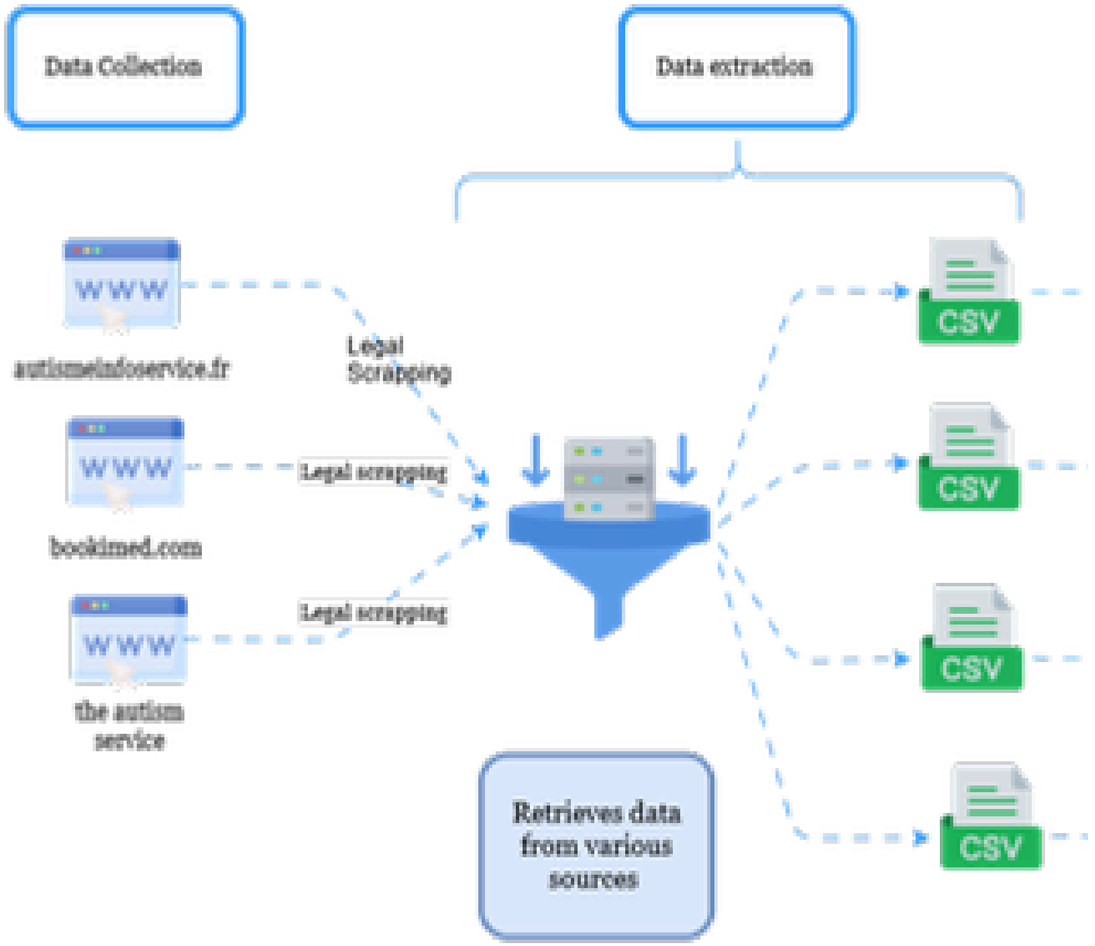
Centers Locations



Source 2: Bookimed (International Clinics Directory)

- Required different extraction strategy due to card-based layouts rather than tabular form, each clinic card parsed individually
- Only few clinics were parsed.

Centers Locations



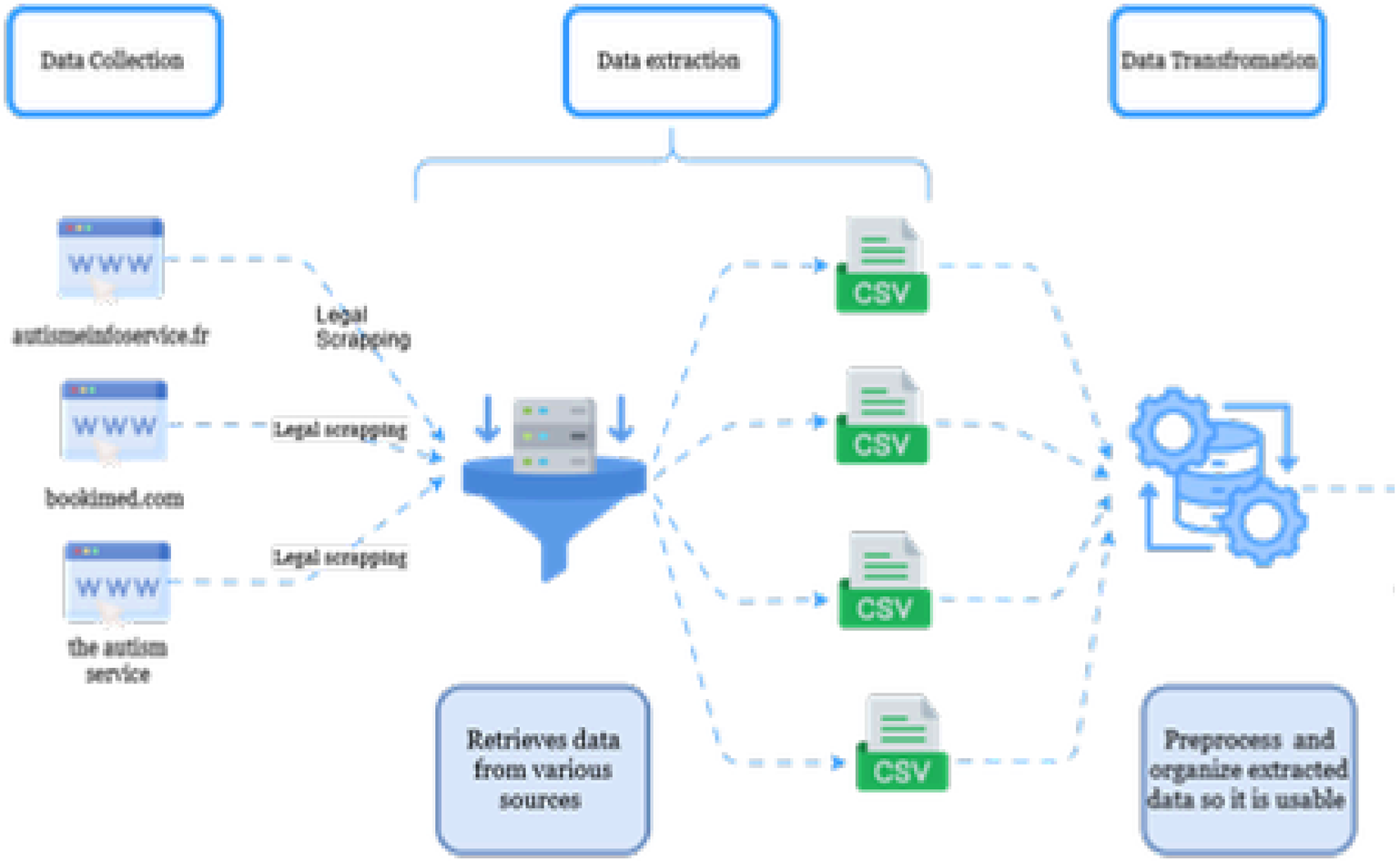
Source 3: The autism service (UK Clinics Directory)

- Public websites listing autism clinics and services.
- Python + Selenium for navigating dynamic web pages.
- Load the clinic listing pages in a browser session.
- Extract text for name, address, phone.
- Handle pagination or “load more” buttons to get all entries.

Overall, the extraction phase required adapting the scraping logic to heterogeneous website structures while dealing with limitations such as:

- incomplete data
- inconsistent HTML markup
- restricted access to some sources.

Centers Locations



Transformation

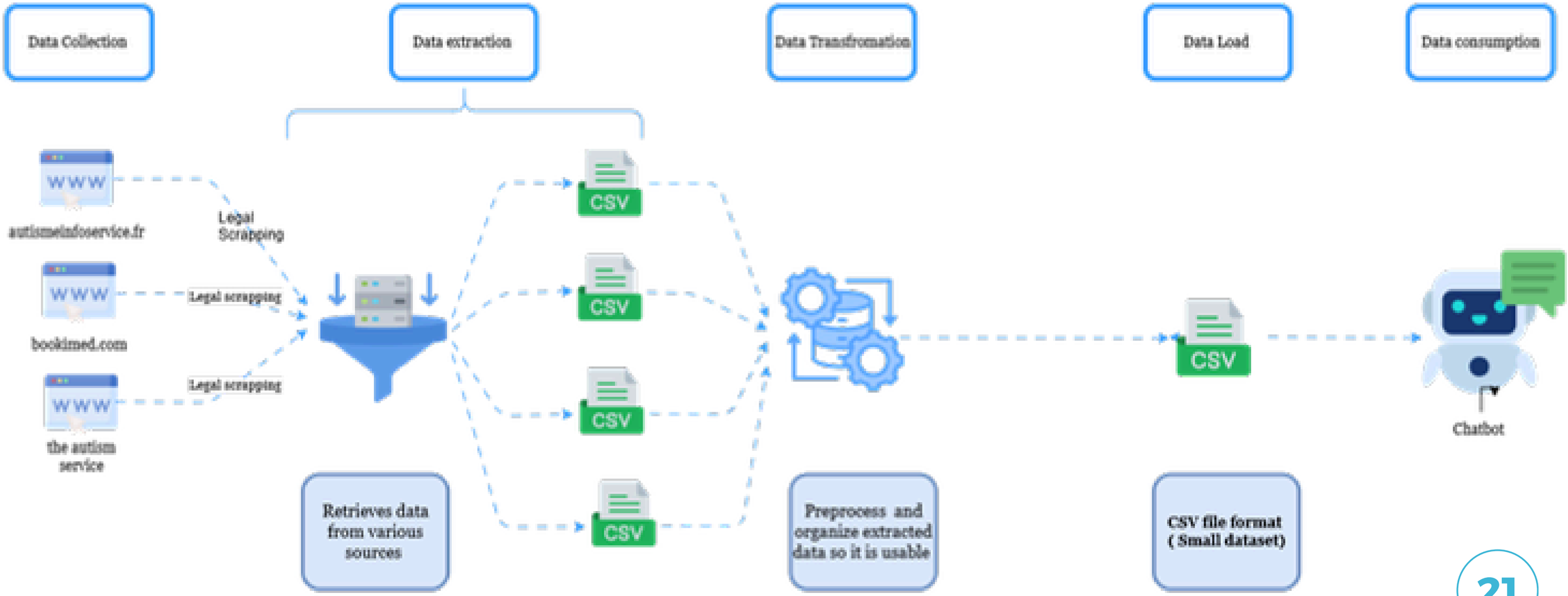
Data Cleaning & Normalization:

- Standardized address formats (split into street number, street name, city, county, postal code).
- Removed redundant columns (e.g., original full address after splitting).

Deduplication:

- Identified and removed duplicate entries to avoid repeated clinic information.

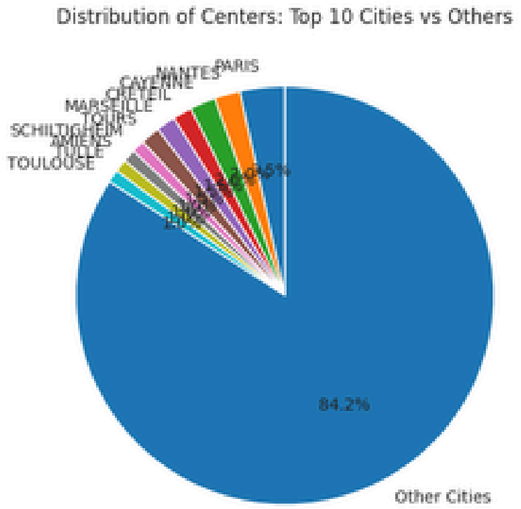
Centers Locations

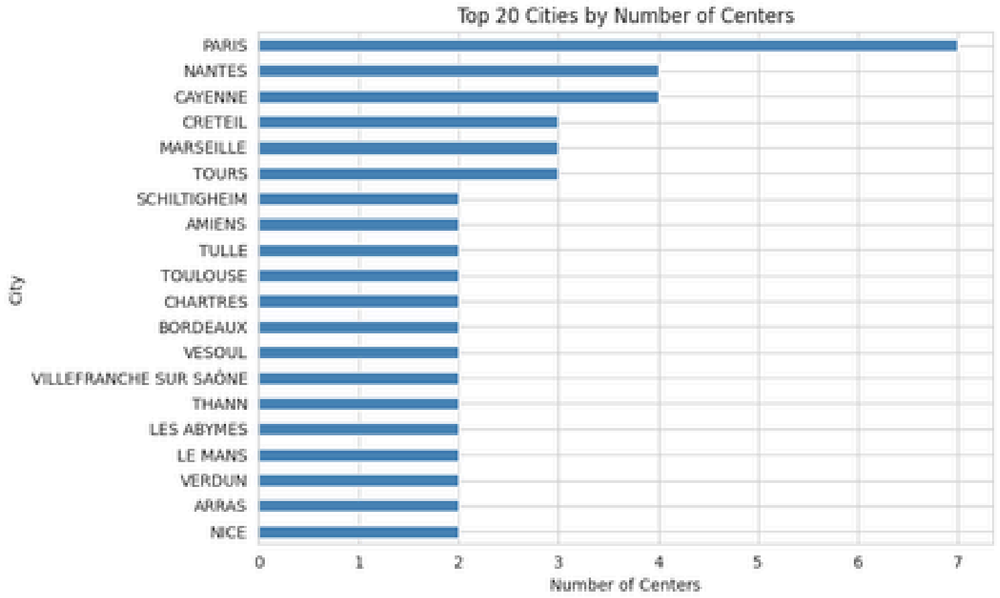


Centers Locations

Loading the datasets

- Given the relatively small size of the datasets and their simple structure, the CSV file format was chosen as the most appropriate solution.
- The cleaned and transformed datasets were exported to CSV files using UTF-8 encoding to preserve special characters. These files can be then used directly for analysis and querying with pandas.

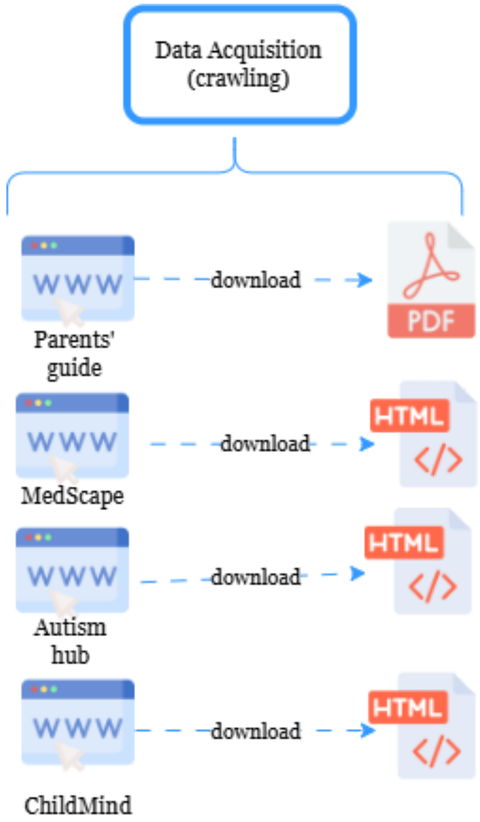




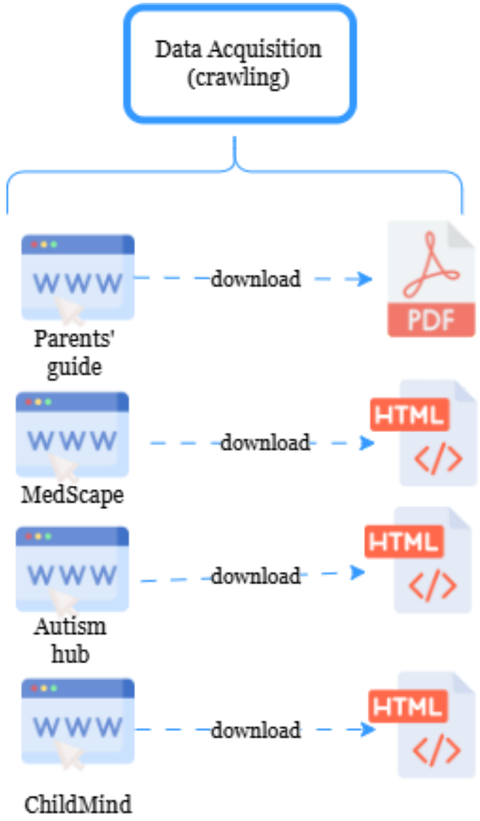
AUTISM CONNECT

Question & Answer (Q&A)

QUESTION & ANSWER (Q&A)



QUESTION & ANSWER (Q&A)

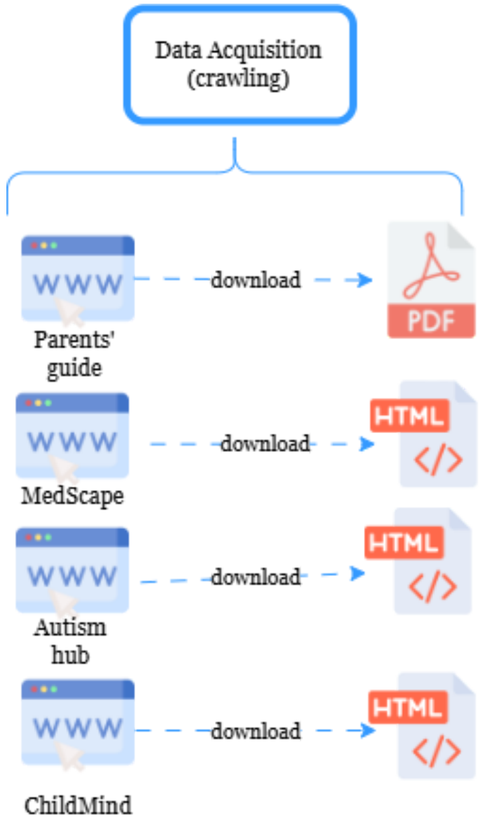


Autism speaks (parent's guide to autism)

- Access requires submitting a form that redirects to a PDF download link
- Form submission returns HTML only; no API or structured endpoint is exposed
- The website indicates that the PDF link is stable and reusable (**Link will open in a new tab/window. Please bookmark the URL for access in the future**)
- Decision: Directly download the PDF, avoiding browser automation (e.g. Selenium)

https://www.autismspeaks.org/tool-kit/parents-guide-autism

QUESTION & ANSWER (Q&A)

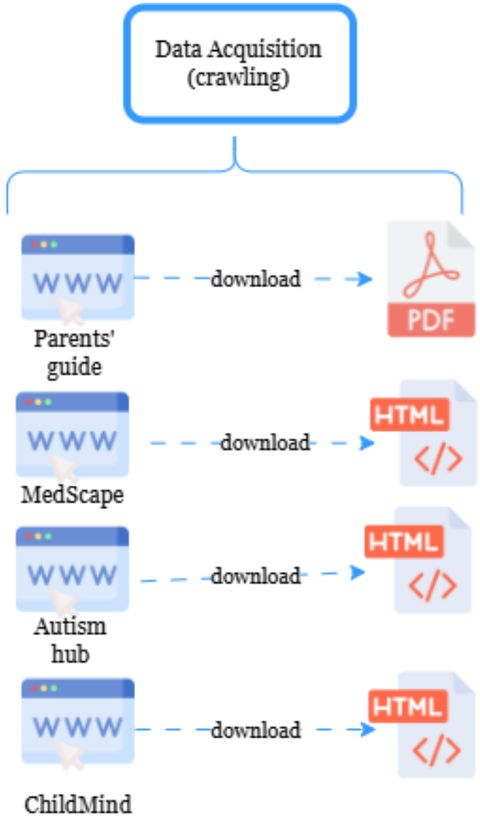


Medscape

- Downloadable article listing questions exist but is disallowed for crawling (robots.txt : **Disallow: /article/*-print$**)
- Decision: Only standard article pages were crawled using scrapy, in compliance with site policies.
- The crawling process started from the root URL and iteratively navigated through subsequent pages by extracting URLs via the clickable “Next” button.

https://emedicine.medscape.com/article/912781-overview

QUESTION & ANSWER (Q&A)

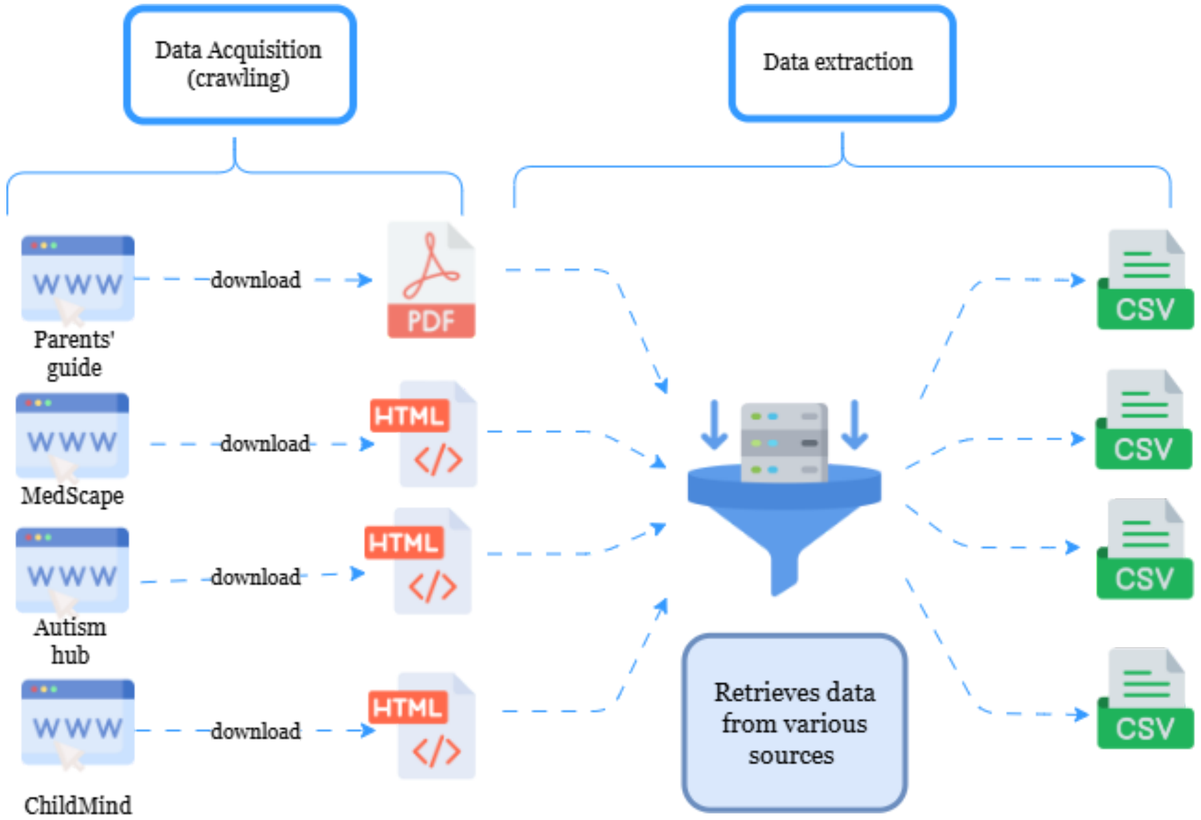


Autism hub & ChildMind

- Content is publicly available as HTML pages
- Decision: Crawled using Scrapy

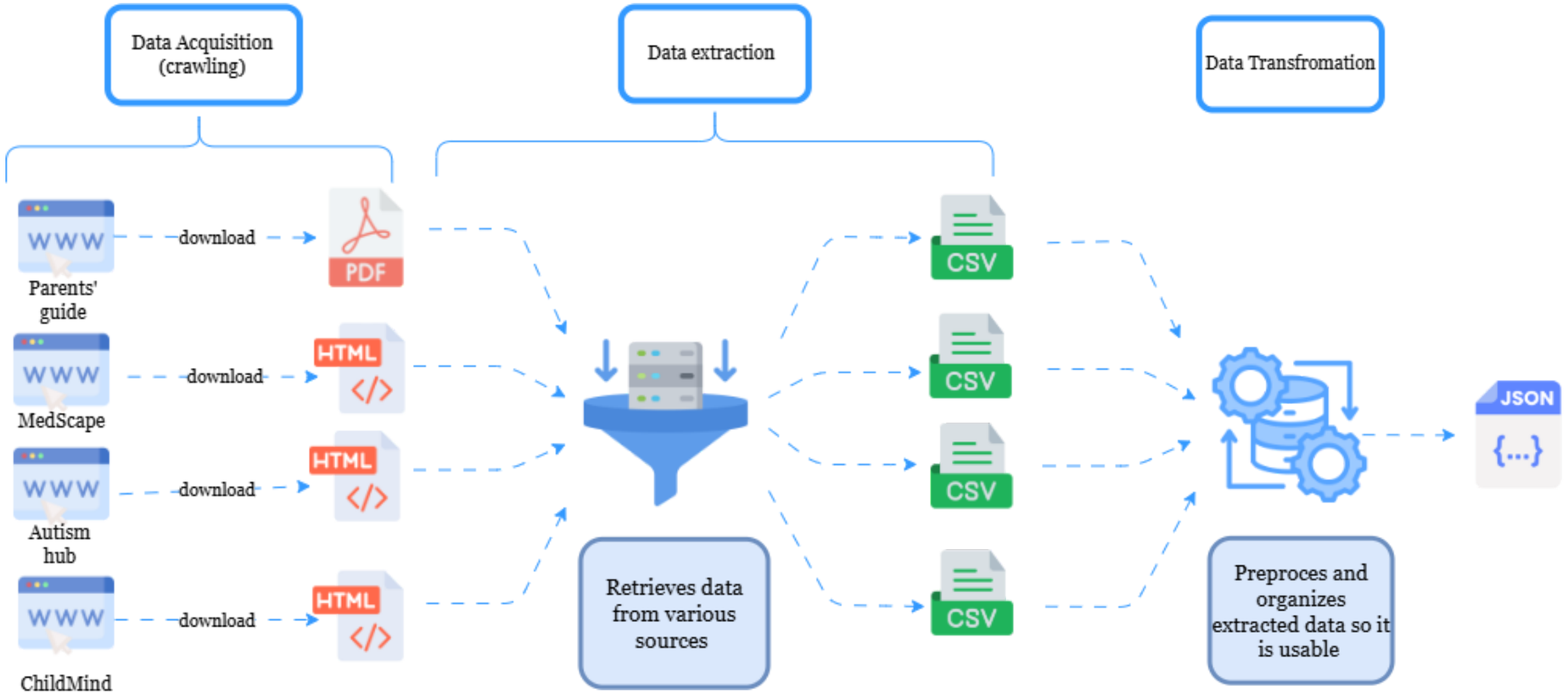
https://autismhub.ie/ask-autism-hub/
https://childmind.org/guide/parents-guide-to-autism/

QUESTION & ANSWER (Q&A)

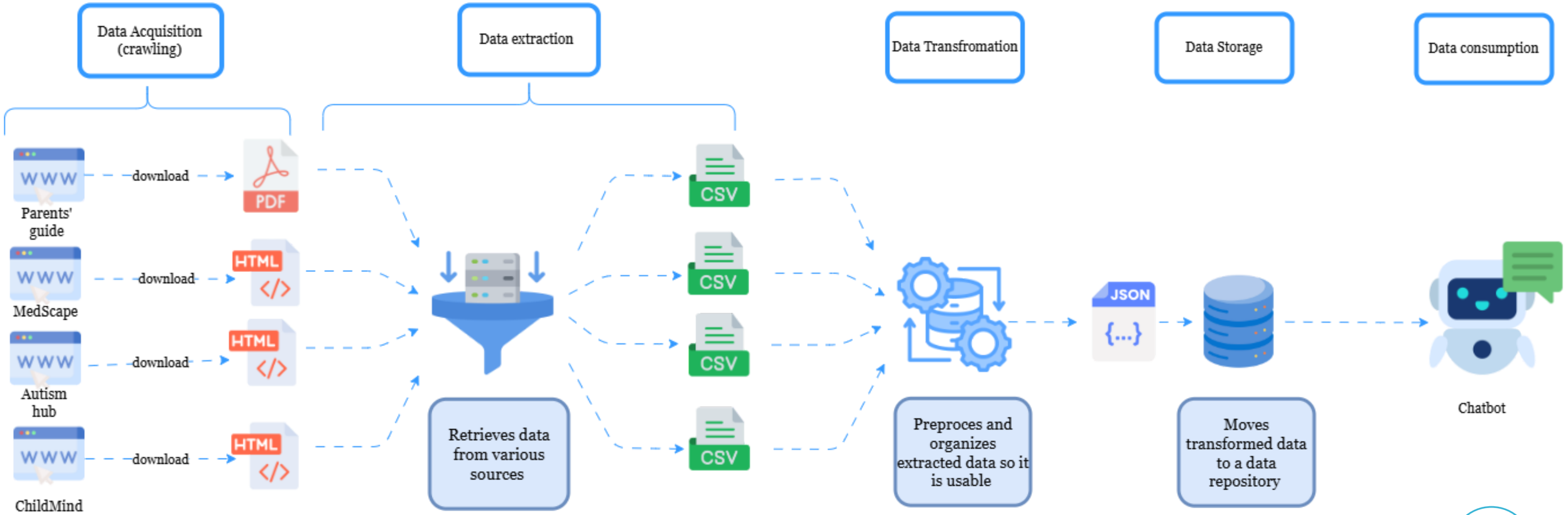


- **PyPDF2** to extract Q&A pairs from the pdf file
- **BeautifulSoup** to extract Q&A pairs from HTML Pages
- For Medscape, questions were extracted manually from the downloaded PDF to respect the robots.txt file while answers were extracted from the html pages with **BeautifulSoup**

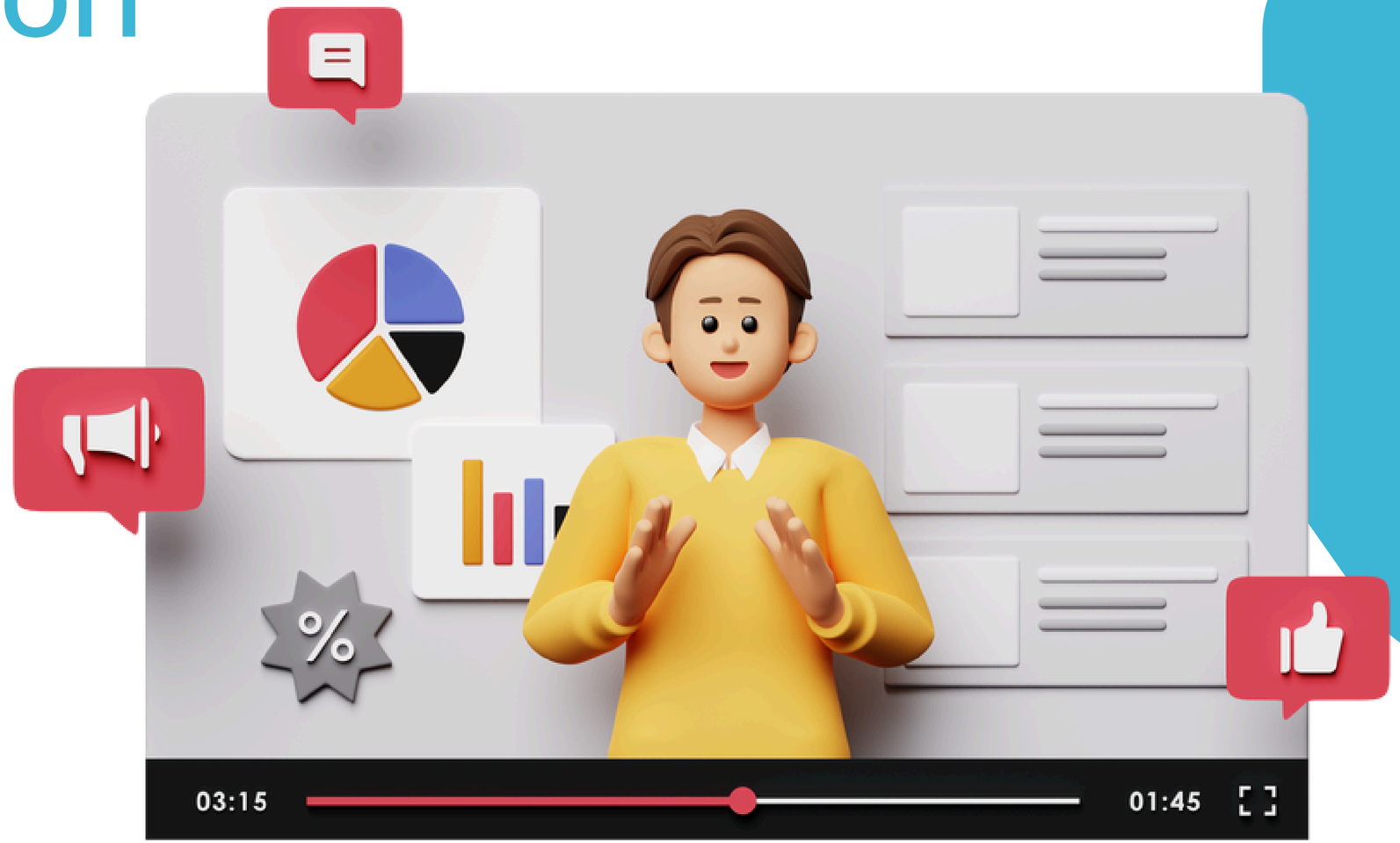
QUESTION & ANSWER (Q&A)



QUESTION & ANSWER (Q&A)



Demonstration



Conclusion

- Built a clean, ethical, and reproducible data pipeline
- Adapted extraction strategies to heterogeneous data sources
- Respected robots.txt and platform constraints
- Prioritized medical reliability and data quality