Collecting & Analising Data for Elissa

Bruxelles, 27/09/2020

Philippe-Fimmers, Francesco Mariottini and Opap's Ditudidi

From Objectives to Modules

OBJECTIVES

1) Collecting data for 10000 buildings.

2) Removing duplicates.

3) Replacing empty information with None.

4) Replacing True/False with "Yes or No.

INDIPENDENT MODULES

a) get_search_results (1)

b) scrap_search_result (1)

c) dataquality (2, 3, 4)

Workflow with CSV creation

Scrapping results solutions

Mansion for sale
€600,000

4 bedrooms | 330 m²
Boulevard de la Dodaine 7 | 1400 — Nivelles

Request your mortgage loan
Immoweb code : 8930262

NEW

photos   map

Log in to see statistics about this property

⚠ Report a mistake   🖨 Print

Overview

🛏 4 bedrooms   🏠 330 m² livable space
🛁 2 bathrooms   ⬚ 390 m² of land

Description

-

- -

General

Available as of	Upon exchange of deeds
Construction year	1924
Building condition	Good
Street facade width	6 m
Facades	3
Covered parking spaces	1
Outdoor parking spaces	2
Virtual visit	Link to the virtual visit

House ⌄

Property sub-types | Reset

☐ Bungalow   ☐ Chalet
☐ Castle   ☐ Farmhouse
☐ Country cottage   ☐ Exceptional property
☐ Apartment block   ☐ Mixed-use building
☐ Town-house   ☐ Mansion
☐ Villa   ☐ Other properties
☐ Country house   ☐ Pavilion

Apartment ⌄

Property sub-types | Reset

☐ Ground floor   ☐ Duplex
☐ Triplex   ☐ Studio
☐ Penthouse   ☐ Loft
☐ Kot   ☐ Service flat

◆

◆

Scrapping results solutions

◆

◆

{property_type}

{page_number}

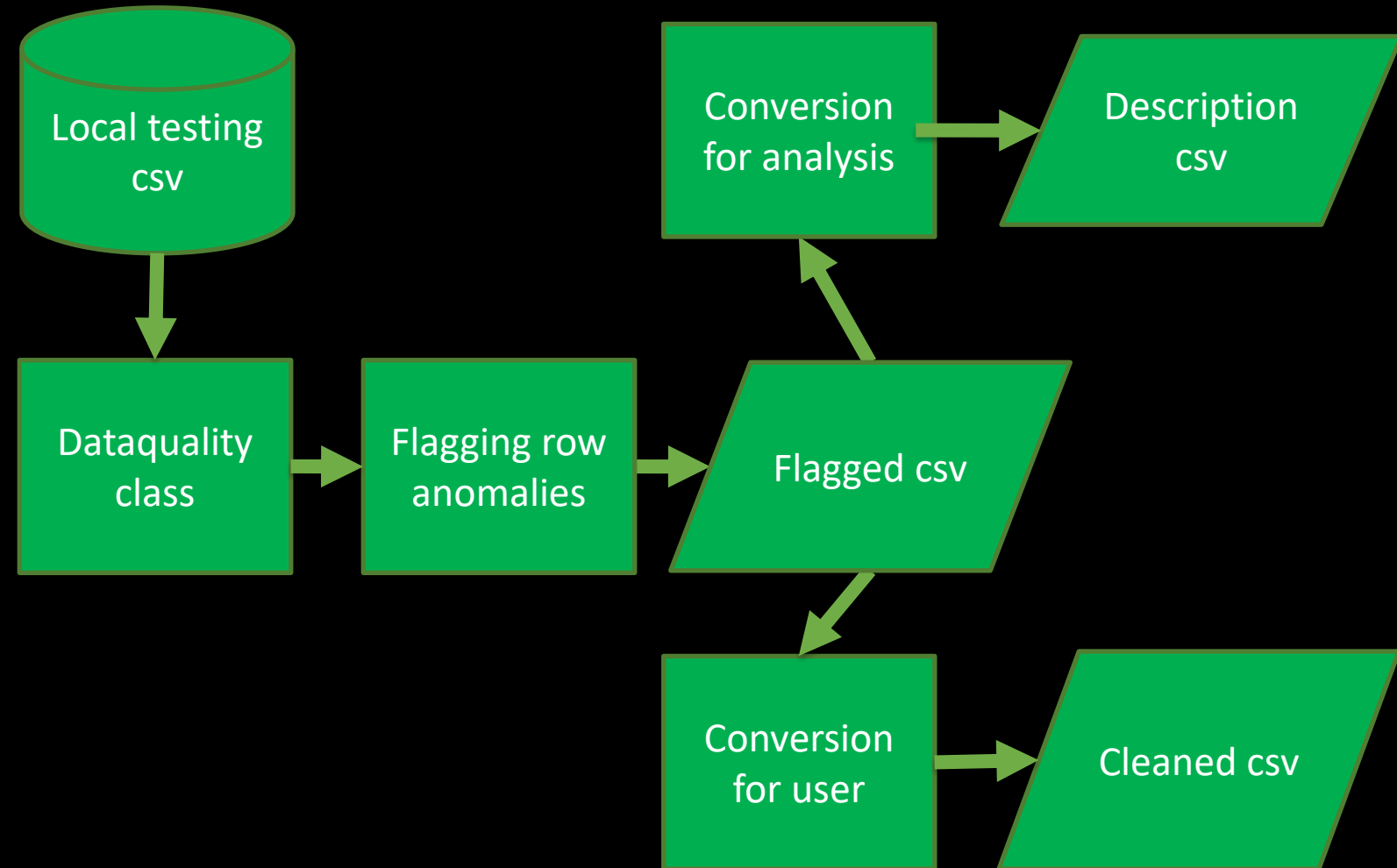# Scrapping results solutions

Scrapping results solutions

- Expected input : Dictionary {url string, is_house bool}

- Expected output : Dictionary {property string, value list}

- scrap_list : dispatches every url from the input list and builds the output dict

- scrap : receives one url and fetches the needed info

Dataquality solutions

CHALLENGES

- Building information not initially available.
- Building information not properly formatted.
- No explicit criteria on duplicates.
- Desired output (str) less effective for analysis.

Local testing csv

Dataquality class

Flagging row anomalies

Flagged csv

Conversion for analysis

Description csv

Conversion for user

Cleaned csv

Lesson learnt

- Clear input & output specifications.
- Internal workflow & deadlines.
- Working packages for parallel testing.
- Previous training for effective coding & support.