

Building a Data Pipeline for Impact Afghanistan

Sarya Baran Kılıç

Laidlaw · Brown University · Impact Afghanistan, Center for Digital Scholarship

Intro

Impact Afghanistan collects publications about Afghanistan from after 2021 and works to find where reporting on the country is missing or hard to access. Doing this at scale requires automation, since no small team can find, organize, and describe this much material by hand. This project built that automation: scrapers for UN and civil society sources, an AI based summarization step, and the fixes needed to get the data working reliably in production.

Background

Impact Afghanistan is connected to the UN's Global Digital Compact and works with Women for Afghan Women (WAW). Its sources range from formal UN records to grassroots first person accounts. That range shaped nearly every decision in building this pipeline.

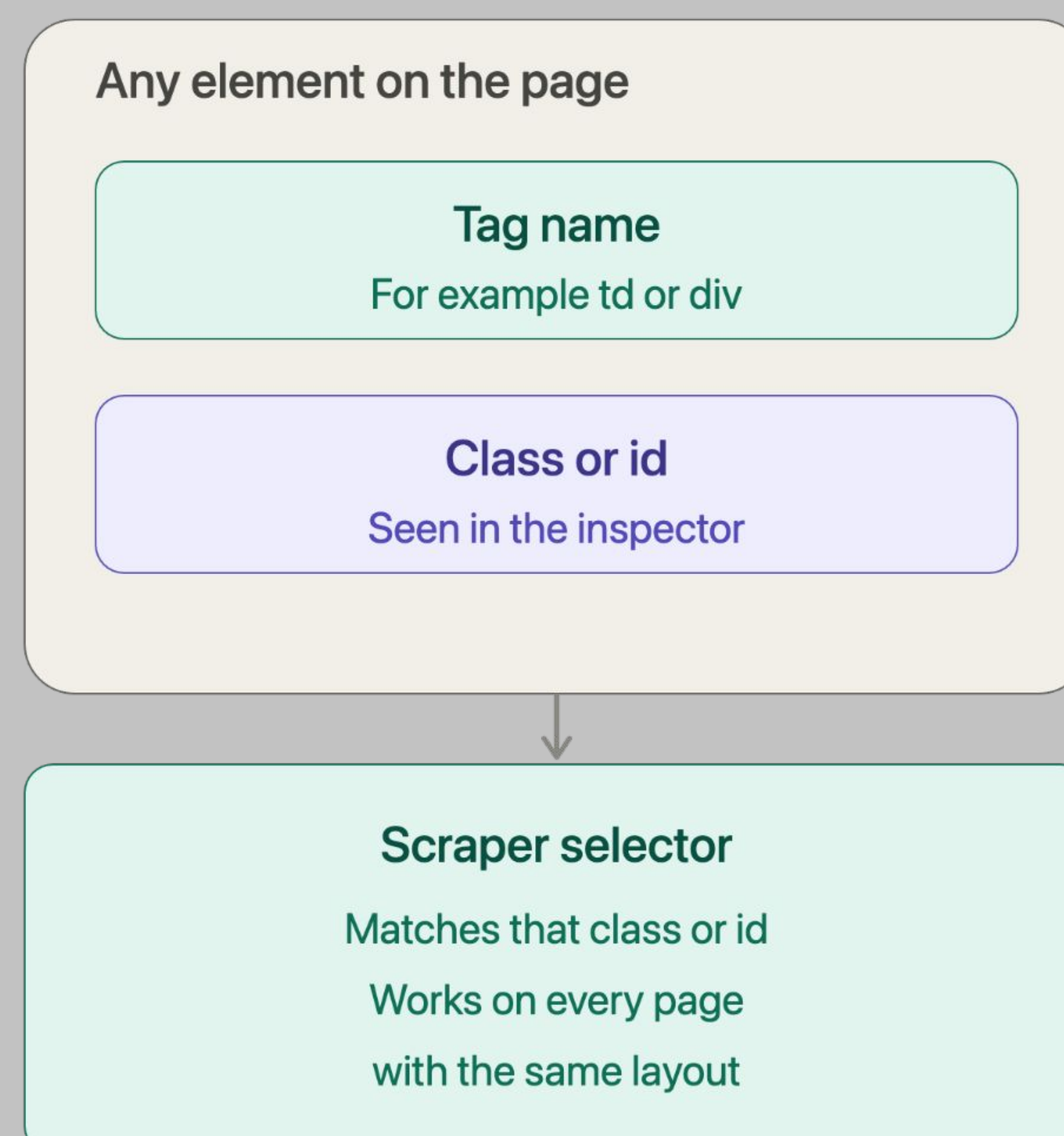
Objectives

The pipeline runs in three stages, shown below: pages are acquired through scraping, each document is enriched with a summary and topic tags, and the results are stored in a database that serves the live site.



Methods

Finding information on a page starts with inspecting it directly, shown below. Once an element's tag or class is identified in the browser, the scraper is written to target that same tag or class across every page of the same type.



Results/Discussion

Challenge 1: The UN Digital Library Blockade

The UN Digital Library indexes documents from every UN body with structured metadata already filled in: title, author, date, and subject, consistent across record types. A single working scraper against this one source would have covered most of what four separate scrapers ended up covering instead. Requests to it were blocked, and no working method was found around that block. This decision point shaped the rest of the pipeline. Without one central source, each UN body's documents had to be collected from its own website, each with its own layout and its own failures.

Challenge 2: Sites That Block Scraping

The same kind of blocking showed up again at the level of individual sites. One source rejected every method tried, so further attempts were stopped once the block was confirmed as intentional and well built. UN Women's pages behaved differently: going straight to a page failed, but clicking into it from another page worked. The scraper now clicks through pages the way a person would, instead of loading them directly.

Future Directions

UN records are built to be read by machines. First person accounts from Afghans are not, and they are the easiest thing for an automated pipeline to miss. Closing real gaps in reporting means dealing with that directly: automating the last step that gets data onto the live site, giving every scraper the same rules for handling missing information, and deciding clearly which blocked sources are worth more effort and which are not.