**Web scraping for CPI**

**Project and Capability Strengthening Plan**

**1. Introduction**

This document sets out a draft curriculum and project plan for the web scraping for CPI project. Sections 2 and 3 provide a brief description of the project and the proposed project outcomes; Section 4 sets out the ways of working; Section 5 outlines the NSO Charter and the commitment expected from NSO participants as well as the Programme Management Unit (PMU) of the UN Regional Hub for Africa; and Section 6 outlines the individual Project Steps & Training Plan for the project.

**2. Description**

This project will explore the possibilities with web scraped prices for use within the CPI. Many learning areas are identified, including the development of Python coding skills, the building web-scrapers, understanding the methodological requirements when considering the use of web scraped data for CPI calculations, assessing the quality of the web scraped data including the coverage and consistency, as well as considerations around the Ethics/Terms and Conditions for web scraping prices. The potential for web scraped prices to inform an early indicator will also be investigated.

The project will be collaborative in nature and will bring NSOs with a mutual interest together to explore this topic, aiding knowledge exchange and skills building across borders.

**3. Project Outcomes**

Participating NSOs may have other objectives and outcomes for this collaborative project, however the core set of outcomes are:

* To understand the extent and range of online prices for CPI, and to assess how useful it is in CPI.
* To understand the methodological considerations for using web scraped Prices for use in CPI calculations.
* To appreciate the ethical considerations prior to embarking on web scraping projects.
* To build skills and application in Python coding and the construction of web-scrapers using appropriate libraries and commands.
* To be able to consider and develop web scrapers for application in other (non-CPI) official statistical outputs.
* To explore the potential for different insights such as early indicators or additional indicators which are difficult to calculate traditionally.
* To develop a pool of ‘new resources’ that can be used for regional cooperation for other NSOs.
* To develop use cases, with associated methodology and code that can be shared via the UN global platform and made available for other NSOs to use in their own web scraping CPI projects.
* To deliver webinars for all African member states to access via the UN Regional Hub for Africa (these will be recorded and made available via the UN Regional Hub for Africa’s dedicated website).

**4. Ways of Working**

A draft outline for how the project will run is shown in the Project Steps & Training Plan (Section 6, Table 1). Ahead of the more formal working and learning laid out in Table 1, there will be:

* A Training Needs Assessment will be used to assess the baseline knowledge and skills of all participants prior to the start of the project. The findings will be used to inform the Project Steps and Training Plan, and hence may undergo some changes before the start of the project. The Needs Assessment will be undertaken remotely, by correspondence. Completion date given is Weds 21st December 2022.
* An introductory session that outlines and agrees this Project and Capability Strengthening Plan, including the scope, timeframe and resource commitment from each NSO. This will also discuss any pre-project work that the training providers will require to be completed prior to the first module taking place. *(2 hours)*
* Development of a ‘Charter’ that the NSO agrees to comply with when participating in the project. This is shared in draft form at Section 5 and will be discussed further for agreement at the introductory session. An updated version will be circulated after that session for NSOs to formally agree.

The aim of the project is to be collaborative and supportive to achieve the outcomes of the project. **The proposed ways of working are:**

* The virtual modules set out in Project Steps & Training Plan (Section 6, Table 1), will be instrumental in moving the participants through the project work. Pre-project work, hands-on training plus homework tasks will be set to ensure the appropriate skills are being built and the project work (application) is progressed between each module.
* The Modules will be virtual and conducted through Microsoft Teams in English
* Each module will have clear outcomes and a plan for achieving the outcomes of the session. These will be shared ahead of each module commencing along with any necessary preparation activities.
* Each module will finish with ‘homework’ or a set of activities for country participants to complete in between each module. These will be to cement the learning of the week in preparation for the next module or undertake research/analysis to inform the next stages of the project.
* In addition to the focussed learning during the module sessions, an ad hoc advice and/or mentoring service will be available for participants. This will allow participants to ask questions or seek advice on the most recent module or homework activities.

**5. NSO Charter**

There is a requirement on participating countries to commit time to the Project Steps & Training Plan (Section 6, Table 1), pre-course work and homework since this is considered fundamental to the capability building requirements plus the delivery of the project.

The following commitments are requested from participants:

* Commitment to the modules as set out in the Project Steps & Training Plan (Section 6, Table 1). Where a participant is unable to attend, for whatever reason, please inform the organisers in good time.
* A commitment to undertake any pre-project work. This will ensure that any appropriate research and/or pre-project decisions can be made in good time to inform the next stage of the project. This will ensure a smooth progression through the project steps. Please submit any pre-project work to the organisers in good time, where requested.
* A commitment to undertake any homework. This will ensure that new learning is applied immediately via project activities.
* A commitment of *one day per week, in addition to the training offered,* to undertake project activities. This is the current estimate for certain members of the project teams (not all) to complete the required project work. This may fluctuate throughout the project and a request is made for all participants to remain flexible to the time commitment required for completion of the project steps

The Programme Management Unit (PMU) of the Africa Regional Hub, will ensure that all participants have access to learning and support throughout the duration of the project through the provision of:

* Hands-on training (virtual) to support individual learning needs and progression of this project.
* Ongoing support in between modules from web scraping and Prices experts. This will be provided in the form of ad hoc advice and/or mentoring.
* A week of face-to-face training and project discussions, in Kigali in March 2023.

The PMU will also undertake a continuous evaluation throughout the duration of the project to ensure that the project outcomes are still being met. Based on the findings, the PMU may at any time adjust the Project Steps & Training Plan (Table 1), to ensure that participants’ learning, and application remains at the core of the project. For example, the PMU may adjust the time between the modules for a better application of learning or may repeat part of a module to better embed principles.

Project Participant NSO: …………………… PMU Member Name:……………………..

Project Participant Name:………………….. PMU Member signature:………………….

Project Participant Signature: ……………..

**Project Steps & Training Plan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Timing** | **Module and description** | **Audience\*** | **Lead** |
| Pre-course work | **Induction**  **19th Jan 2023** | * What e-commerce shops are available in country? Are they online only or also have physical shops? * Do the e-commerce shops transport products across the country or only in selected areas? Are there significant differences in transport rates for the different areas? * Can the NSO access any reliable data, for example, for magnitude of sales or are there any tax reports for those e-commerce companies? Any info here will be used to “weight” the data * Provide a check list of questions for them to research and submit report * Have a website that they want to use to scrape for real in module 5 (week 7) | CPI experts within NSO | Luigi Palumbo |
| Pre-course environment set-up | **Induction**  **19th Jan 2023** | Issue list of software requirements for installation onto laptops & answer any immediate questions:   * Establish the training environment for the Coders and understand any possible limitations with laptops, Python, etc * Issue guidance for downloading Python to their laptops (we will need to know which operating system they are using) * All coders to install Python with associated libraries onto their laptops, and any other required tools, such as Jupyter notebooks * Installation of a standard open-source browser and driver for browser automation [I suggest Chromium and chromium-driver, those are open source and available on all OS]. Include Selenium IDE browser extension * Installation of GitHub Desktop * Test and debug prior to the start of the training | Coders | Dominik Dabrowski |
| Pre-course environment set-up | **Environment Surgery**  **Thu 2nd Feb** | Meeting/Surgery to check pre-course environment. |  | Dominik Dabrowski |
| **Capability Strengthening & Project development** | | | | |
| Module 1  (*2 hours)* | **Wed 8th Feb 2023** | **Introduction to web-scraping Prices**   * Outline the basic requirements for web-scraping * The requirements that must be in place (eg ethical permissions and a sample / web addresses) * Introduce concept and requirements for a strategy for price scraping: * What areas of CPI do they want to improve? * Which product groups, etc? * What other information do you need? * Introduction to Methodology considerations – coverage, volume, adherence and integration with other data streams * Leave the team with homework to undertake research and principles to consider when developing a strategy (using a strategy template) * Activities to enable and inform **Module 5** to go ahead | CPI experts within NSO | Luigi Palumbo |
| Module 2  (2 hours) | **Wed 15th Feb 2023** | **Agreeing web-scaping strategies**   * Country teams to quickly present results from the pre-course work * Review country strategies with the aim of finalising a roadmap at the end of the session * Definition of a minimum viable pilot for each country team as outcome from the cooperation project, including:   + Target data source(s)   + Scraping schedule, data pipeline, and monitoring   + Final data product   + Project documentation | CPI experts within NSO | Luigi Palumbo |
| Module 3  *(3 x 3-hour sessions)* | **W/c 20th Feb**  *Mon 20 Feb*  *Wed 22 Feb*  *Fri 24 Feb*  *13:00 – 16:00 CET* | **Introduction to Python**   * Github – platform for sharing and storing (Luigi – 20th Feb) * The basic foundations of python (Dominik) * Leave with exercises to cement learning * Provide an ad-hoc mentoring service over the next three weeks for questions or assistance | Coders | Dominik Dabrowski + Luigi Palumbo |
| Module 4  *(3 x 3-hour sessions)* | **W/c 13th Mar**  *Mon 13 Mar*  *Wed 15 Mar*  *Fri 17 Mar*  *13:00 – 16:00 CET* | **Designing a general web scraper in Python**   * Applying python code to web scraping – using different examples (including prices) * Scrape dummy websites – data extraction & assessment of the website’s structure * Data Cleaning * Leave with exercises to cement learning * Provide an ad-hoc mentoring service over the next 2 weeks for questions or assistance | Coders | Dominik Dabrowski |
| **In-person week of Capability strengthening: 27th – 31st March 2023** | | | | |
| Module 5 | **W/c 27th Mar**  *NISR Training Centre, Kigali* | **Web Scraping for CPI (Part 1)**   * Using the strategy developed pre-course and in Module 1 & 2 – walk through steps to web scrape prices * Start scraping their website of choice (selected during pre-course work) | Coders and CPI experts | Luigi Palumbo & Dominik Dabrowski |
| Module 6 | **W/c 27th Mar**  *NISR Training Centre, Kigali* | **Web Scraping for CPI (Part 2)**   * Assess scraping results and refine code etc to improve scraped pool, what other info is required. * Discuss automation, data pipelines, expected finished products * Discuss Methodology | Coders and CPI experts | Luigi Palumbo & Dominik Dabrowski |
|  | **W/c 27th Mar**  *NISR Training Centre, Kigali* | Other potential topics (to be determined):   * Refine web scraping and ‘pool’ of prices data to inform conclusions * What conclusions can we draw from the scraping of prices? * What do the prices tell us about inflation/early indicators outside of official statistics? * Report writing and outcomes from this Project * What other topics can we use web scraping on to complement official statistics or other data gaps/policy topics? * Introduction to automation/pipelines – applying automation techniques to your CPI data collection or other survey processes at the NSO | Coders and CPI experts | Luigi Palumbo & Dominik Dabrowski |