17 January 2025

English

United Nations Group of Experts on Geographical Names
2025 session
New York, 28 April – 2 May 2025
Item 5 (a) of the provisional agenda [*]
Technical expertise: Names collection, office treatment, features beyond a single sovereignty and international cooperation

The Utilization of Artificial Intelligence for Geographical Name Data Web Scraping

Summary [**]

Web scraping is a widely used technique for extracting data from the internet, enabling collecting reliable and up-to-date information in large volumes. However, creating web scraping scripts is often complex and time-intensive, mainly when dealing with sophisticated and dynamic web pages. This paper explores the application of artificial intelligence (AI) in automating web scraping script development, specifically for geographical name data extraction in Indonesia.

The methodology involved utilizing ChatGPT, a generative AI model, to automate the scriptwriting process. The process began with identifying relevant data sources and the specific elements to be extracted, such as geographical names, coordinates, and related metadata. Challenges associated with dynamic content and complex data structures were addressed using generative AI to produce Python scripts incorporating libraries such as Selenium for interacting with dynamic elements and BeautifulSoup for parsing HTML. ChatGPT was instrumental in compiling, refining, and optimizing these scripts. Iterative testing and adjustments were conducted to ensure functionality across diverse web pages.

The results demonstrate that utilizing AI significantly enhances efficiency in script development without compromising quality. AI effectively manages dynamic content and data hidden behind JavaScript. However, this paper also highlights the necessity of manual validation to ensure script reliability and data accuracy, particularly for complex structures. Additionally, ethical and legal considerations were emphasized, underscoring the need to comply with privacy policies, copyright laws, and website owner permissions. The responsible use of scraped data is essential to maintaining fairness and adherence to regulatory standards.

[*] GEGN.2/2025/1.
[**] Prepared by Kiki Nur Putra, Yusnita Permana, and Nafisa Andika Putri, Indonesia, Geospatial Information Agency (BIG). The report will be available under document symbol GEGN.2/2025/X/CRP.X, in the language of submission only, at https://unstats.un.org/unsd/ungegn/sessions/4th_session_2025/