**Technical Specification**

**Publication Content Collection System for Info Sapiens Website**

**1. Purpose**

Development of a Python-based automated system for collecting and organizing publications from the specified website's publication section, including both web page content and associated PDF reports.

Website: <https://www.sapiens.com.ua/en/publications>

**2. Functional Requirements**

**2.1. Data Collection**

* The system must access and process all publications from the specified URL
* For each publication, the system must:
  + Extract and save the main text content from the web page
  + Download the full PDF report when available
  + Preserve all tables, images, and formatted content from the web page
  + Maintain the original formatting structure of the content

**2.2. File Management**

* Convert web page content into PDF format
* Save two types of PDF files when applicable:
  + Short version: containing the web page content
  + Full version: the downloaded complete report
* File naming convention:
  + Format: [publication\_date]*[publication\_title]*[version].pdf
  + Where:
    - publication\_date: date of the publication
    - publication\_title: clean version of the publication title
    - version: "short" or "full"

**2.3. Content Processing**

* Handle both English and Ukrainian language content
* Preserve all data tables and visual elements
* Maintain proper formatting of text and structural elements

**3. Output Requirements**

* All files must be saved in a structured directory
* Each publication should result in either:
  + One PDF file (short version) if no full report is available
  + Two PDF files (short and full versions) if a full report is available
* Generated PDFs must be searchable and allow text selection

**4. Quality Requirements**

* Complete capture of all publication content
* Accurate preservation of data and formatting
* Proper handling of special characters in filenames
* Reliable processing of multilingual content

**5. Additional Considerations**

* The system should respect the website's terms of service and access policies
* Processing should handle various potential errors (network issues, unavailable content, etc.)
* The system should be able to process publications regardless of their size or content type

**6. Deliverables**

The completed system should provide:

* Automated collection of all publications
* Conversion of web content to PDF format
* Download of available full reports
* Organized storage of all collected materials
* Clear documentation of the implemented solution

**7. Success Criteria**

* All publications are successfully collected and stored
* PDF files are properly named and organized
* Content is accurately preserved in the generated PDFs
* Both short and full versions are correctly managed
* System reliably handles all types of publication content