Problem Definition

With the rapid evolution of the internet, online content is frequently modified, moved, or deleted, leading to broken links and loss of valuable information. Users who rely on web pages for research, reference, or personal archiving often face challenges in preserving content for future access. Existing solutions may be complex, require subscriptions, or fail to provide a convenient offline storage format. There is a need for a minimalistic and efficient tool that allows users to archive web pages seamlessly and store them in a structured, portable format.

Project Objective

A minimal tool for downloading and archiving a website which is aimed to be used on personal devices where the webpage/website will be stored as **ePUB**. It is aware of network constraints namely bandwidth and total data available to work with, and has interface to ensure user does not download more than they are prepared to, which is not the focus of most other similar tools.

Proposed Plan of Work

Literature Review

- Archiving techniques
- Convert for file formats ie html to xhtml
- Authoring *ePUB* file

Requirement Analysis

- Site scraping
- Show page content size
- Converting to epub

Testing

- Testing bounded site scraping
- Conversion to epub for small sites

Evaluation

- Evaluating the scraper on large site and profiling
- Mapping web routes to epub chapters/section

Documention & Reporting

- Creating a man page
- Guides for installation

Methodology

- Fetching specified webpages and its assets
- Recursively fetch all subsequent page if needed.
- Remap all the routes to point pages stored in file system.
- Sanitize and convert HTML to XHTML.
- Creating manifest file and table of content, etc for ePUB.
- Archive all the file finally in ePUB.

Technology

- **Go**
  - Has good concurrency pattern
  - Has feature rich standard library particularly for http client and server
- **XML**

    - It is foundation of epub
- **HTML-Tidy**
    - HTML-Tidy corrects and clean up HTML content by fixing markup errors such as mismatched, mis-nested, and missing tags; missing end “/” tags; missing quotations; and many, many more discrepant conditions, and serves as an HTML pretty printer.
- **Pandoc**
    - A popular inter-document conversiong tool

Functional Specification (Deliverables)

- Fetching and storing webpage/website recurisively.
- Convertion of HTML to XHTML.
- Arhciving as ePUB.

Project Scope

A minimal web archiving tool that enables users to fetch and store webpages or entire websites recursively. The tool will ensure content preservation by converting HTML to well-structured XHTML, maintaining compatibility and consistency across different devices. Additionally, the archived content will be bundled into an EPUB format, providing a portable and easily accessible offline reading experience. The project will focus on efficiency, simplicity, and usability, ensuring a seamless process for users to save and organize web content for future reference.