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

Testing

Evaluation

Documention & Reporting

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
- XML
- TidyHtml
- Pandoc

Functional Specification (Deliverables)

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

Project Scope