

LU13a - Crawler Aufbau Teil 1

Damit alle relevanten Seiten von DokuWiki (<https://wiki.bzz.ch>) heruntergeladen werden können, nutzen wir eine Kombination aus Collector, Downloader und Orchestrator.

WikiPageDownloaderService

Der WikiPageDownloaderService lädt den Inhalt einer Seite via XML-RPC herunter. Das hat mitunter zur Folge, dass kein HTML-Dokument, sondern direkt der DokuWiki-Markup-Code zurückgeschickt wird.

Dieser Code ist einiges schlanker als das entsprechende HTML und macht das Interpretieren einfacher. Das HTML-Dokument enthält wiederkehrende Elemente wie Header und Footer, die nicht relevant sind. Ebenfalls ist viel Logik und Styling ebenfalls im HTML-Dokument enthalten. Das nachfolgende Beispiel zeigt den Unterschied.

[illegible]

From:

<https://wiki.bzz.ch/> - **BZZ - Modulwiki**

Permanent link:

<https://wiki.bzz.ch/de/modul/ffit/3-jahr/java/learningunits/lu13/a?rev=1764630542>

Last update: **2025/12/02 00:09**

