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.


Rendered HTML

WikiPageCollectorService

Der `WikiPageCollectorService` sammelt die URLs von den Seiten, die heruntergeladen werden sollen.

Der Benutzer soll einen Unterordner beziehungsweise einen „Namespace“ angeben können und sämtliche darunterliegenden Seiten sollen rekursiv gesammelt werden.

Leider gibt es dazu keine geeignete und funktionierende XML-RPC-Funktion. Theoretisch könnte man alle Seiten als HTML-Dokumente runterladen und dann sämtliche links (<a href=…) speichern.

In unserem Fall nutzen

Daher nutzen wir dieselbe Funktionen, die auch vom Wiki verwendet wird, um Unterseiten aufzulisten.

```
<nspages . -h1  -exclude -simpleList -textPages="">
```

`Ajax` https://wiki.bzz.ch/start?idx=

WikiCrawlerPipelineService

Orchestriert wird das Ganze Filter