01. Rozszerzenia serwera dLibra

Wprowadzenie

Mechanizm rozszerzeń serwera dLibra bazuje na bibliotece Java Plugin Framework (JPF). Podstawowym elementem w tym mechanizmie jest plik opisujący plugin JPF:

```
<?xml version="1.0" ?>
<!DOCTYPE plugin PUBLIC "-//JPF//Java Plug-in Manifest 0.7" "http://jpf.sourceforge.net/plugin_0_7.dtd">
<plugin id="pl.psnc.dlibra.content" version="$Revision: 1.2 $"
        vendor="PSNC">
        <extension-point id="extraction.TextualContentExtractor">
                <parameter-def id="class" />
                <parameter-def id="order" />
        </extension-point>
</plugin>
```

Powyższy plik definiuje tylko jeden punkt rozszerzenia serwera opisane poniżej.

> **Uwaga**
>
> Interfejsy programistyczne wyszczególnione w poniższych opisach znajdują się w bibliotece programistycznej `dlibra-server-extension-api`

.

Punkt rozszerzenia **extraction.TextualContentExtractor**

Zestaw rozszerzeń `extraction.TextualContentExtractor` służy do ekstrakcji tekstu z plików z treścią publikacji. Dzięki temu możliwe jest indeksowanie treści publikacji niezależnie od jej formatu. Dla tekstowych formatów dokumentów, takich jak na przykład HTML, dostęp do tekstu jest niemal natychmiastowy. W przypadku innych formatów pliki muszą być w odpowiedni sposób przygotowane, żeby ekstrakcja tekstu była możliwa. Tylko wówczas rozszerzenia będą w stanie taki tekst uzyskać i przekazać do indeksacji.

Opisywane rozszerzenie przyjmuje dwa parametry:

- `class` - nazwa klasy, która implementuje interfejs programistyczny tego rozszerzenia
- `order` - parametr odpowiadający za kolejność wybierania danego rozszerzenia (pozwala to na ustalenie, które z rozszerzeń będzie użyte w przypadku, gdy jest więcej niż jedno rozszerzenie obsługujące dany format treści)

Interfejs programistyczny (język Java) dla tego rozszerzenia to `pl.psnc.dlibra.content.extraction.TextualContentExtractor`. Bardziej szczegółowe informacje na temat jego działania znajdują się w dokumentacji programistycznej (JavaDocs).

Serwer dLibry dostarczany jest z preinstalowanym zestawem rozszerzeń tego typu. Należą do nich:

- Ekstrakcja tekstu z prostych formatów takich jak CHM, HTML, RTF, TXT (`dlibra-server-extension-tce-basic`).
- Ekstrakcja tekstu z formatu DjVu (`dlibra-server-extension-tce-djvu`).
- Ekstrakcja tekstu z formatów obsługiwanych przez zewnętrzny mechanizm LIUS (`dlibra-server-extension-tce-lius`).
- Ekstrakcja tekstu z formatu PDF - (`dlibra-server-extension-tce-pdf`).

Rozszerzenia te opisano pokrótce poniżej.

Rozszerzenie basic

Rozszerzenie to posiada zestaw klas implementujących interfejs `pl.psnc.dlibra.content.extraction.TextualContentExtractor` i pozwalających wyciągać tekst z plików w następujących formatach:

- **CHM** - klasa `pl.psnc.dlibra.content.extraction.CHMTextualContentExtractor`
- **HTML** - klasa `pl.psnc.dlibra.content.extraction.HTMLTextualContentExtractor`
- **RTF** - klasa `pl.psnc.dlibra.content.extraction.RTFTextualContentExtractor`
- **TXT** - klasa `pl.psnc.dlibra.content.extraction.TXTTextualContentExtractor`

Rozszerzenie djvu

Rozszerzenie to pozwala wyciągnąć tekst z warstwy tekstowej plików w formacie DjVu (jeśli posiadają taką warstwę). Jest to realizowane przez klasę `pl.psnc.dlibra.content.extraction.DjVuTextualContentExtractor`

Rozszerzenie lius

Rozszerzenie to wykorzystuje zewnętrzną bibliotekę LIUS (Lucene Index Update and Search), która umożliwia m.in. wyciąganie tekstu z plików w następujących formatach: MsWord, MsExcel, MsPowerPoint, RTF, PDF, XML, HTML, TXT, OpenOffice, ZIP, MP3, VCard, Latex i JavaBeans. Klasa rozszerzenia, która realizuje wyciąganie tekstu z wyżej wymienionych formatów plików to `pl.psnc.dlibra.content.extraction.LIUSTextualContentExtractor`.

Rozszerzenie pdf

Wyciąganie tekstu z plików w formacie PDF oparte jest o zewnętrzną bibliotekę PDFBox. Klasa implementująca interfejs `pl.psnc.dlibra.content.extraction.TextualContentExtractor` i realizująca ekstrakcję tekstu z tego typu plików to `pl.psnc.dlibra.content.extraction.PDFTextualContentExtractor`.