[EN] 01. The Extensions of the dLibra Server

Introduction

he extension mechanism of the dLibra server is based on the Java Plugin Framework (JPF) library. The basic element of that mechanism is the JPF plugin description file:

```
<?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>
```

The file shown above only defines one server extension point described below.

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

.

The extraction.TextualContentExtractor Extension Point

The `extraction.TextualContentExtractor` extension set is for extracting text from files with publication content. The extension makes it possible to index publication content regardless of its format. In the case of text document formats, such as HTML, text can be accessed almost instantly. In the case of other formats, files must be prepared to make text extraction possible. Only then will extensions be able to extract the text and pass it on to be indexed.

That extension has two parameters:

- `class` – the name of the class which implements the programming interface of the extension, and
- `order` – the parameter which is responsible for extension selection order (which makes it possible to determine which extension will be used when more than one extension can handle the given content format).

The programming interface (Java language) for that extension is `pl.psnc.dlibra.content.extraction.TextualContentExtractor`. For more information about it, see the programming documentation (JavaDocs).

The dLibra system is provided with a pre-installed set of extensions of that type. The set includes:

- text extraction from simple formats, such as CHM, HTML, RTF, or TXT (`dlibra-server-extension-tce-basic`),
- text extraction from the DjVu format (`dlibra-server-extension-tce-djvu`),
- text extraction from formats supported by an external mechanism, LIUS (`dlibra-server-extension-tce-lius`), and
- text extraction from the PDF format (`dlibra-server-extension-tce-pdf`).

Those extensions are briefly described below.

The Basic Extension

The basic extension has a set of classes which implement interface `pl.psnc.dlibra.content.extraction.TextualContentExtractor` and make it possible to extract text from files in the following formats:

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

The DjVu extension

The DjVu extension makes it possible to extract text from the text layer of files in the DjVu format (if they have such a layer). That task is done by class `pl.psnc.dlibra.content.extraction.DjVuTextualContentExtractor`

The LIUS Extension

The LIUS extension makes use of the (external) LIUS (Lucene Index Update and Search) library which allows, among other things, text extraction from files with the following formats: MsWord, MsExcel, MsPowerPoint, RTF, PDF, XML, HTML, TXT, OpenOffice, ZIP, MP3, VCard, Latex, and JavaBeans. The extension class which extracts text from those file formats is `pl.psnc.dlibra.content.extraction.LIUSTextualContentExtractor`.

The PDF extension

Extracting text from files in the PDF format is based on the (external) PDFBox library. The class which implements interface `pl.psnc.dlibra.content.extraction.TextualContentExtractor` i and extracts text from files of that type is `pl.psnc.dlibra.content.extraction.PDFTextualContentExtractor`.