ExtractingUpdateProcessor

ExtractingUpdateProcessor

⚠ Currently under development in SOLR-1763

Introduction

The ExtractingUpdateProcessor is an Update Processor capable of extracting text out of rich documents such as PDFs and MS Office documents and more. It is based on Apache Tika which has support for more than 30 document formats. The processor is shipped in the `solr-extraction` contrib module, bundled together with ExtractingRequestHandler.

ExtractingUpdateProcessor does the same job as ExtractingRequestHandler, namely extracting text from rich documents. But using it as an UpdateProcessor has several benefits over the RequestHandler approach:

- Extract text from multiple binary attachments in the same Solr document
- Better control of which fields to write the output and metadata to
- Use with any RequestHandler, such as XML, CSV, Binary (SolrJ), DIH etc (since all these support the UpdateChain)
- Do more complex integrations, like an UpdateChain which reads a file reference from the document, then fetches the document from external storage before extraction

Configuration

The UpdateRequestProcessor is configured in solrconfig.xml, and supports many parameters. All parameters listed may also be overridded on the update request itself. A minimal configuration will read input from a binary field named `stream_content` and the file name from field `stream_name` and output extracted data to fields `title` and `body`:

```
<processor class="org.apache.solr.update.processor.ExtractingUpdateProcessorFactory" />
```

NOTE: The processor supports the `defaults/appends/invariants` concept for its config. However, it is also possible to skip this level and configure the parameters directly underneath the `<processor>` tag.

Below follows a list of each configuration parameters and their meaning:

⚠ TBD

a

Bla bla

Value: true/false

Default: true

Examples

Override input and output fields

```
<processor class="org.apache.solr.update.processor.ExtractingUpdateProcessorFactory" >
  <str name="in.content.field">binary_content</str>
  <str name="in.filename.field">filename</str>
  <str name="out.title.field">title_en</str>
  <str name="out.body.field">description_en</str>
  <str name="out.mimetype.field">mimetype</str>
</processor>
```

Resources

- Apache Tika
- SOLR-1763
- ExtractingRequestHandler