Tabular Files Extractor

The *Tabular Files Extractor* gets a tabular file (a comma/tab separated file) from the content stream in the Job and converts it into XML, using the Metadata Mapper to map each column to an XML Element. It then will submit each row of the XML file as a separate subJob.

On this page:

Configuration

Element	Type	Default	Description
separator	string	'comma'(,)	The character used to separate each column on the tabular file. It can be any character, 'comma'(,) or 'tab'(\t). Defaults to ','.
headersOnFirstRow	boolean	false	Whether to use the first row data as column names or not. If false, columns will be named column1, column2, ..., in the metadata mapping.
ignoreQuotes	boolean	false	When false it will treat unescaped (backslash escaping) quotes as delimiters for literal text (separators will be considered as normal characters). If true it will include the quotes as normal characters
metadataMap	Map	null	field names to map each column. See Metadata Mapper for more info.
branches		None	The configuration of the pipeline to publish to. See below.

Branch Configuration

The Tabular SubJob Extractor publishes documents using the branch manager. It publishes using the events configured above. You must therefore include <branches> for these events in the configuration to publish to a pipeline within a pipeline manager. See Branch Handler for more details.

Element	Type	Description
branches/branch/@event	String	The event to configure. Should always be "onSubJob".
branches/branch/@pipelineManager	string	The URL of the pipeline manager to publish to. Can be relative.
branches/branch/@pipeline	string	The name of the pipeline to publish to.

Example Configuration

```
<component name="TabularSubJobExtractor" subType="default" factoryName="aspire-tabular-files">
   <branches>
     <branch event="onSubJob" pipelineManager="." pipeline="subJobsPipeline" />
   </branches>
   <separator>tab</separator>
   <headersOnFirstRow>true</headersOnFirstRow>
 </component>
```

Example Configuration with No Headers on Input File

```
<component name="TabularSubJobExtractor" subType="default" factoryName="aspire-tabular-files">
   <branches>
     <branch event="onSubJob" pipelineManager="subjobPipelineManager" />
   </branches>
   <separator>tab</separator>
   <headersOnFirstRow>false</headersOnFirstRow>
   <metadataMap>
     <map from="column0" to="term"/>
     <map from="column1" to="frequency"/>
     <map from="column2" to="type"/>
     <map from="column3" to="id"/>
   </metadataMap>
 </component>
```

Example Use Within A Pipeline

```
  <pipeline name="process-feedOne-test">
   <stages>
     <stage component="fetchUrl" />
     <stage component="TabularSubJobExtractor" />
   </stages>
 </pipeline>
```

Example

In the following example suppose that there's a file called "file:test.txt" which contains the following:

```
first        second        third
data1        data2        data3
```

Further suppose that "file:test.txt" is read by the Fetch URL stage. Once executing the Tabular SubJob Extractor, each subJob will contain a row of the original document, which in this case, is only one row:

```
 <doc>
    <parent><fetchUrl>./testdata/com.accenture.aspire.components/testdata/testcommaseparated.csv</fetchUrl><
/parent>
    <subDocId>test.txt-0</subDocId>
    <extension source="TabularSubJobExtractor">
       <field name="first">data1</field>
       <field name="second">data2</field>
       <field name="third">data3</field>
    </extension>
  </doc>
```