Metadata Splitter

The *Metadata Splitter* stage parses fields with delimited lists and creates multiple <val> tags. These nested tags are easier to manipulate with XSLT for later processing (e.g. post-xml)

Metadata Splitter	
Factory Name	com.searchtechnologies.aspire:aspire-tools
subType	splitter
Inputs	AspireObject with metadata text content with delimiters that need to be split into separate XML tags.
Outputs	AspireObject

Configuration

Element	Type	Default	Description
delimeter	string	;	Specify the default delimeter string to use to split the metadata elements. For example: <delimiter>,</delimiter>
xPath	string	none	Specify xPath element(s) in the AspireObject e.g. /doc/category. All elements matched by the xPath will be split. *Note: <xPath> statements can be specified.*
xPath /@delimiter	string	none	Each <xPath> can take a delimiter to specify how to split elements matched by that particular xPath.
tag	string	none	Specify an XML tag in the AspireObject e.g. "category" to split. Only splits the first matching tag in the <doc>. Runs substantially faster than the <xPath> command. *Note: <tag> statements can be specified.*
tag /@delimiter	string	none	Each <tag> can take a delimiter to specify how to split that particular tag.

Notes:

- **Warning**: The entire content (including nested XML tags) of any tag matched by the instructions supplied to the splitter above will be deleted and replaced with <val> tags.
- Content split will automatically be trim()ed (leading and trailing spaces removed).

Sample Configuration

```
<component name="splitter" factoryName="aspire-tools" subtype="splitter">
  <!-- xPath: match anywhere and can match multiple elements, all are split -->
  <xPath>//geographicArea</xPath>
  <xPath>//category</xPath>
  <xPath delimiter=\":\">//searchKeywords</xPath>

  <!-- tag: matches only the first matching element at the top level, but runs faster -->
  <tag>subCategory</tag>

  <!-- Specify the default delimiter -->
  <delimter>;</delimter>
</config>
```

Example

The following example uses the configuration specified above.

Before:

```
 <doc>
   <fetchURL>+www.oilandgaspurchaser.com</fetchURL>
   <feederLabel>CrawlSinglePage</feederLabel>
   <category source="CCDMeta/category">Data;  ;Companies;Gas/LNG;Europe;  ;Crude Petroleum and Natural Gas<
/category>
   <geographicArea source="CCDMeta/geographicArea">All Aspermont Oil and Gas domains;All UK domains<
/geographicArea>
   <urltitle source="CCDMeta/urltitle"/>
   <acronym source="CCDMeta/acronym"/>
   <urldescription source="CCDMeta/urldescription"/>
   <urlcomments source="CCDMeta/urlcomments"/>
   <category source="CCDMeta/category">Data;  ;Companies;Gas/LNG;Europe;  ;Crude Petroleum and Natural Gas<
/category>
   <startURL source="CCDMeta/startURL">www.oilandgaspurchaser.com</startURL>
   <subCategory>Organizations;Surface Mining;Mineral Processing;Engineering;Underground Mining;Metals &
Minerals</subCategory>
   <keywords>
       <searchKeywords source="CCDMeta/searchKeywords1">080624: ERROR: BADDNS</searchKeywords>
   </keywords>
 </doc>
```

After:

```
 <doc>
   <fetchURL>+www.oilandgaspurchaser.com</fetchURL>
   <feederLabel>CrawlSinglePage</feederLabel>
   <category source="CCDMeta/category">
       <val>Data</val>
       <val>Companies</val>
       <val>Gas/LNG</val>
       <val>Europe</val>
       <val>Crude Petroleum and Natural Gas</val>
   </category>
   <geographicArea source="CCDMeta/geographicArea">
       <val>All Aspermont Oil and Gas domains</val>
       <val>All UK domains</val>
   </geographicArea>
   <urltitle source="CCDMeta/urltitle"/>
   <acronym source="CCDMeta/acronym"/>
   <urldescription source="CCDMeta/urldescription"/>
   <urlcomments source="CCDMeta/urlcomments"/>
   <category source="CCDMeta/category">
       <val>Data</val>
       <val>Companies</val>
       <val>Gas/LNG</val>
       <val>Europe</val>
       <val>Crude Petroleum and Natural Gas</val>
   </category>
   <startURL source="CCDMeta/startURL">www.oilandgaspurchaser.com</startURL>
   <subCategory>
       <val>Organizations</val>
       <val>Surface Mining</val>
       <val>Mineral Processing</val>
       <val>Engineering</val>
       <val>Underground Mining</val>
       <val>Metals & Minerals</val>
   </subCategory>
   <keywords>
       <searchKeywords source="CCDMeta/searchKeywords1">
           <val>080624</val>
           <val>ERROR</val>
           <val>BADDNS</val>
       </searchKeywords>
   </keywords>
 </doc>
```