XMLParser Plugin

Xml parser (parse-xml) is **configurable plugin**. It use **XPath** and **namespaces** to do the mapping between XML elements and Lucene fields.

Informations :

1- Copy "xmlparser-conf.xml" to the nutch/conf dir

2- To index your custom XML file, you have to modify the "xmlparser-conf.xml". This parser uses namespaces and XPATH to parse XML content The config file do the mapping between the XML noeds (using XPATH) and lucene field. Example : <field name="dctitle" xpath="//dc:title" type="Text" boost=" 1.4" />

3- The xmlIndexerProperties encapsulate a set of fields associated to a namespace. If the namespace is found in the xml document, the fields represented by the namespace will be indexed. Example :

```
<xmlIndexerProperties type="filePerDocument" namespace=" http://purl.org/dc/elements/1.1/">
  <field name="dctitle" xpath="//dc:title" type="Text" boost=" 1.4" />
  <field name="dccreator" xpath="//*[name()='dc:creator']" type="keyword" boost=" 1.0" />
</xmlIndexerProperties>
```

4- It is possible to define a default namespace that will be applied when the parser didn't find any namespace in the document or when the namespace found in the xml document doesn't match with the namespace defined in the xmlIndexerProperties. Example :

```
<xmlIndexerProperties type="filePerDocument" namespace="default">
  <field name="xmlcontent" xpath="//*" type="Unstored" boost="1.0" />
</xmlIndexerProperties>
```

You can download parse-xml plugin in : http://issues.apache.org/jira/browse/NUTCH-185

Contact :rida.benjelloun@doculibre.com