**<?xml version="1.0" encoding="UTF-8" ?>**

*<!--*

*Licensed to the Apache Software Foundation (ASF) under one or*

*more contributor license agreements. See the NOTICE file*

*distributed with this work for additional information regarding*

*copyright ownership. The ASF licenses this file to You under the*

*Apache License, Version 2.0 (the "License"); you may not use*

*this file except in compliance with the License. You may obtain*

*a copy of the License at*

*http://www.apache.org/licenses/LICENSE-2.0 Unless required by*

*applicable law or agreed to in writing, software distributed*

*under the License is distributed on an "AS IS" BASIS, WITHOUT*

*WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.*

*See the License for the specific language governing permissions*

*and limitations under the License.*

*-->*

*<!--*

*Description: This document contains Solr 3.1 schema definition to*

*be used with Solr integration currently build into Nutch. See*

*https://issues.apache.org/jira/browse/NUTCH-442*

*https://issues.apache.org/jira/browse/NUTCH-699*

*https://issues.apache.org/jira/browse/NUTCH-994*

*https://issues.apache.org/jira/browse/NUTCH-997*

*https://issues.apache.org/jira/browse/NUTCH-1058*

*and*

*http://svn.apache.org/viewvc/lucene/dev/branches/branch\_3x/solr/*

*example/solr/conf/schema.xml?view=markup*

*for more info.*

*-->*

<schema name="nutch" version="1.5">

<types>

<fieldType name="text" class="solr.TextField" positionIncrementGap="100" autoGeneratePhraseQueries="true">

<analyzer type="index">

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="1" catenateNumbers="1" catenateAll="0" splitOnCaseChange="1"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory" />

<filter class="solr.SnowballPorterFilterFactory" language="English" />

</analyzer>

<analyzer type="query">

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="0" catenateNumbers="0" catenateAll="0" splitOnCaseChange="1"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory" />

<filter class="solr.SnowballPorterFilterFactory" language="English" />

</analyzer>

</fieldType>

<fieldType name="string" class="solr.StrField" sortMissingLast="true"

omitNorms="true"/>

<fieldType name="long" class="solr.TrieLongField" precisionStep="0"

omitNorms="true" positionIncrementGap="0"/>

<fieldType name="int" class="solr.TrieIntField" precisionStep="0" omitNorms="true" positionIncrementGap="0"/>

<fieldType name="float" class="solr.TrieFloatField" precisionStep="0"

omitNorms="true" positionIncrementGap="0"/>

*<!-- boolean type: "true" or "false" -->*

<fieldType name="boolean" class="solr.BoolField" sortMissingLast="true" omitNorms="true"/>

*<!--Binary data type. The data should be sent/retrieved in as Base64 encoded Strings -->*

<fieldtype name="binary" class="solr.BinaryField"/>

<fieldType name="double" class="solr.TrieDoubleField" precisionStep="0" omitNorms="true" positionIncrementGap="0"/>

<fieldType name="tint" class="solr.TrieIntField" precisionStep="8" omitNorms="true" positionIncrementGap="0"/>

<fieldType name="tfloat" class="solr.TrieFloatField" precisionStep="8" omitNorms="true" positionIncrementGap="0"/>

<fieldType name="tlong" class="solr.TrieLongField" precisionStep="8" omitNorms="true" positionIncrementGap="0"/>

<fieldType name="tdouble" class="solr.TrieDoubleField" precisionStep="8" omitNorms="true" positionIncrementGap="0"/>

<fieldType name="tdate" class="solr.TrieDateField" omitNorms="true" precisionStep="6" positionIncrementGap="0"/>

*<!-- A general unstemmed text field - good if one does not know the language of the field -->*

<fieldType name="textgen" class="solr.TextField" positionIncrementGap="100">

<analyzer type="index">

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="1" catenateNumbers="1" catenateAll="0" splitOnCaseChange="0"/>

<filter class="solr.LowerCaseFilterFactory"/>

</analyzer>

<analyzer type="query">

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="0" catenateNumbers="0" catenateAll="0" splitOnCaseChange="0"/>

<filter class="solr.LowerCaseFilterFactory"/>

</analyzer>

</fieldType>

<fieldType name="random" class="solr.RandomSortField" indexed="true" />

<fieldType name="date" class="solr.TrieDateField" precisionStep="0"

omitNorms="true" positionIncrementGap="0"/>

<fieldType name="fullText" class="solr.TextField" positionIncrementGap="100" autoGeneratePhraseQueries="true">

<analyzer type="index">

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="1" catenateNumbers="1" catenateAll="0" splitOnCaseChange="1"/>

<charFilter class="solr.MappingCharFilterFactory" mapping="mapping-FoldToASCII.txt"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory" />

<filter class="solr.SnowballPorterFilterFactory" language="English" />

</analyzer>

<analyzer type="query">

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="0" catenateNumbers="0" catenateAll="0" splitOnCaseChange="1"/>

<charFilter class="solr.MappingCharFilterFactory" mapping="mapping-FoldToASCII.txt" />

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory" />

<filter class="solr.SnowballPorterFilterFactory" language="English" />

</analyzer>

</fieldType>

*<!-- For Spellcheck -->*

<fieldType name="textSpell" class="solr.TextField" positionIncrementGap="100">

<analyzer type="index">

<tokenizer class="solr.StandardTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory"/>

</analyzer>

<analyzer type="query">

<tokenizer class="solr.StandardTokenizerFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" enablePositionIncrements="true" />

<filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory"/>

</analyzer>

</fieldType>

*<!-- lowercases the entire field value, keeping it as a single token. -->*

<fieldType name="lowercase" class="solr.TextField" positionIncrementGap="100">

<analyzer>

<tokenizer class="solr.KeywordTokenizerFactory"/>

<filter class="solr.LowerCaseFilterFactory" />

</analyzer>

</fieldType>

<fieldType name="url" class="solr.TextField"

positionIncrementGap="100">

<analyzer>

<tokenizer class="solr.StandardTokenizerFactory"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.WordDelimiterFilterFactory"

generateWordParts="1" generateNumberParts="1"/>

</analyzer>

</fieldType>

*<!-- Field Defination Suggesting single words -->*

<fieldType class="solr.TextField" name="text\_auto" positionIncrementGap="100">

<analyzer>

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1" catenateWords="1" catenateNumbers="1" catenateAll="0" splitOnCaseChange="1"/>

<filter class="solr.LowerCaseFilterFactory"/>

</analyzer>

</fieldType>

*<!-- Field Defination Suggesting phrase -->*

*<!--*

*<fieldType class="solr.TextField" name="text\_auto">*

*<analyzer>*

*<tokenizer class="solr.KeywordTokenizerFactory"/>*

*<filter class="solr.LowerCaseFilterFactory"/>*

*</analyzer>*

*</fieldType>*

*-->*

<fieldType name="autocomplete" class="solr.TextField" positionIncrementGap="100">

<analyzer>

<tokenizer class="solr.StandardTokenizerFactory"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt"/>

<filter class="solr.ShingleFilterFactory" maxShingleSize="2" outputUnigrams="false"/>

<filter class="solr.RemoveDuplicatesTokenFilterFactory"/>

</analyzer>

</fieldType>

</types>

<fields>

<field name="id" type="string" stored="true" indexed="true" required="true" multiValued="false" />

*<!-- Spellcheck field -->*

<field name="spell" type="textSpell" indexed="true" stored="false" required="false" multiValued="true" />

*<!-- For Sorting purpose -->*

<field name="titleSort" type="lowercase" indexed="true" stored="false" required="false" multiValued="true"/>

*<!-- core fields -->*

<field name="segment" type="string" stored="true" indexed="false"/>

<field name="digest" type="string" stored="true" indexed="false"/>

<field name="boost" type="float" stored="true" indexed="false"/>

*<!-- fields for index-basic plugin -->*

<field name="host" type="url" stored="false" indexed="true"/>

<field name="site" type="string" stored="true" indexed="true"/>

<field name="url" type="url" stored="true" indexed="true"/>

<field name="content" type="text" stored="true" indexed="true"/>

<field name="title" type="text" stored="true" indexed="true"/>

*<!-- <field name="metatag.title" type="text" stored="true" indexed="true" multiValued="true"/> -->*

<field name="cache" type="string" stored="true" indexed="false"/>

<field name="tstamp" type="date" stored="true" indexed="true"/>

<field name="\_version\_" type="long" indexed="true" stored="true"/>

*<!-- fields for index-anchor plugin -->*

<field name="anchor" type="string" stored="true" indexed="true"

multiValued="true"/>

*<!-- Field for auto suggestion -->*

<field name="name\_autocomplete" type="text\_auto" indexed="true" stored="true" multiValued="false" />

<field name="ac-terms" type="autocomplete" indexed="true" stored="false" multiValued="true" omitNorms="true" omitTermFreqAndPositions="false" />

*<!-- fields for index-more plugin -->*

<field name="type" type="string" stored="true" indexed="true"

multiValued="true"/>

<field name="contentLength" type="long" stored="true"

indexed="false"/>

<field name="lastModified" type="date" stored="true"

indexed="false"/>

<field name="date" type="date" stored="true" indexed="true"/>

*<!-- fields for languageidentifier plugin -->*

<field name="lang" type="string" stored="true" indexed="true"/>

*<!-- KS -->*

<field name="metatag.description" type="text" stored="true" indexed="true"/>

<field name="description" type="text" stored="true" indexed="true"/>

<field name="metatag.keywords" type="text" stored="true" indexed="true"/>

<field name="keywords" type="text" stored="true" indexed="true"/>

*<!-- fields for subcollection plugin -->*

<field name="subcollection" type="string" stored="true"

indexed="true" multiValued="true"/>

*<!-- fields for feed plugin (tag is also used by microformats-reltag)-->*

<field name="author" type="string" stored="true" indexed="true"/>

<field name="tag" type="string" stored="true" indexed="true" multiValued="true"/>

<field name="feed" type="string" stored="true" indexed="true"/>

<field name="publishedDate" type="date" stored="true"

indexed="true"/>

<field name="updatedDate" type="date" stored="true"

indexed="true"/>

*<!-- fields for creativecommons plugin -->*

<field name="cc" type="string" stored="true" indexed="true"

multiValued="true"/>

*<!-- This field has been added to implement the default search configuration -->*

<field name="text" type="fullText" indexed="true" stored="false" required="false" multiValued="true" />

*<!-- For Extract Handler -->*

*<!-- copyField commands copy one field to another at the time a document*

*is added to the index. It's used either to index the same field differently,*

*or to add multiple fields to the same field for easier/faster searching. -->*

<copyField source="anchor" dest="text"/>

<copyField source="title" dest="text"/>

<copyField source="content" dest="text"/>

*<!-- <copyField source="description" dest="text"/>-->*

*<!-- <copyField source="keywords" dest="text"/> -->*

*<!-- For Spell Check -->*

<copyField source="anchor" dest="spell"/>

<copyField source="title" dest="spell"/>

*<!-- <copyField source="content" dest="spell"/> commented on 26112015 -->*

<copyField source="keywords" dest="spell"/>

<copyField source="description" dest="spell"/>

*<!-- Copy Field for auto suggestion -->*

<copyField source="title" dest="name\_autocomplete" />

<copyField source="title" dest="ac-terms"/>

*<!-- For Sorting purpose -->*

*<!-- Dynamic field definitions. If a field name is not found, dynamicFields*

*will be used if the name matches any of the patterns.*

*RESTRICTION: the glob-like pattern in the name attribute must have*

*a "\*" only at the start or the end.*

*EXAMPLE: name="\*\_i" will match any field ending in \_i (like myid\_i, z\_i)*

*Longer patterns will be matched first. if equal size patterns*

*both match, the first appearing in the schema will be used. -->*

<dynamicField name="\*\_i" type="int" indexed="true" stored="true"/>

<dynamicField name="\*\_s" type="string" indexed="true" stored="true"/>

<dynamicField name="\*\_l" type="long" indexed="true" stored="true"/>

<dynamicField name="\*\_t" type="text" indexed="true" stored="true"/>

<dynamicField name="\*\_b" type="boolean" indexed="true" stored="true"/>

<dynamicField name="\*\_f" type="float" indexed="true" stored="true"/>

<dynamicField name="\*\_d" type="double" indexed="true" stored="true"/>

<dynamicField name="\*\_dt" type="date" indexed="true" stored="true"/>

*<!-- some trie-coded dynamic fields for faster range queries -->*

<dynamicField name="\*\_ti" type="tint" indexed="true" stored="true"/>

<dynamicField name="\*\_tl" type="tlong" indexed="true" stored="true"/>

<dynamicField name="\*\_tf" type="tfloat" indexed="true" stored="true"/>

<dynamicField name="\*\_td" type="tdouble" indexed="true" stored="true"/>

<dynamicField name="\*\_tdt" type="tdate" indexed="true" stored="true"/>

<dynamicField name="attr\_\*" type="textgen" indexed="true" stored="true" multiValued="true"/>

<dynamicField name="random\_\*" type="random" />

</fields>

<uniqueKey>id</uniqueKey>

<defaultSearchField>text</defaultSearchField>

<solrQueryParser defaultOperator="OR"/>

<similarity class="org.apache.lucene.search.similarities.DefaultSimilarity"/>

</schema>