IndexStructure

The Index Structure

The index structure formed after indexing is shown below :

Field Name	Stored	Index	Plugin/Class	Comment	version	
					1.x	2.x
id	YES	Indexed, Un-Tokenized	IndexerMapReduce/IndexUtil	**URL** used as **ID** to update and delete documents	X	X
boost	YES	Not Indexed	various scoring plugins	Adds a **score** value field to a particular document. This is allocated based upon its importance within the webgraph.	?	?
digest	YES	Not Indexed	org.apache.nutch.indexer. IndexerMapReduce.java	Adds a **message digest** field to a document. Can be MD5 over content and headers or more sophisticated text profile of the content.	?	?
lang	YES	Un-Tokenized	language-identifier	Add a **lang**, language field to a document.	?	?
segment	YES	Not Indexed	org.apache.nutch.indexer. IndexerMapReduce.java	Adds the originating **segment** field to the document, used to identify the most recent segment in which this document was fetched.	?	?
tstamp	YES	Tokenized	index-basic	Adds a **timestamp** field of the most recent time this document was fetched	?	?
cc: license	YES	Indexed, Tokenized	creativecommons	Adds the entire license as **cc:license=xxx** and **attributes** extracted of the license url	?	?
cc:meta	YES	Indexed, Tokenized	creativecommons	Adds the license location as **cc:meta=xxx**	?	?
cc:type	YES	Indexed,Tokenized	creativecommons	Adds the work type as **cc:type=xxx**	?	?
anchor	NO	Tokenized	index-anchor	Indexing filter that indexes all inbound **anchor text** for a document.	?	?
title	YES	Tokenized	index-basic	Adds basic searchable **title field** to a document. Also indexed by index-more	?	?
host	NO	Tokenized	index-basic	Adds basic searchable **hostname field** to a document.	?	?
url	YES	Tokenized	index-basic	Adds basic searchable **URL field** to a document. May differ from "id" in case the page is the redirect target.	?	?
content	NO	Tokenized	index-basic	Adds basic searchable **content field** to a document	?	?
lastModified	NO	Indexed, Un-Tokenized	index-more	Adds some time related meta info in the form of **last-modified** if present.	?	?
date	NO	Indexed, Un-Tokenized	index-more	Index date as last-modified, or, if that's not present, uses fetch time.	?	?
contentLength	NO	Indexed, Un-Tokenized	index-more	⚠️ NEEDS COMMENT ⚠️	?	?
type	NO	Indexed, Un-Tokenized	index-more	Adds contentType, primaryType, subType (all mime-types)	?	?
primaryType	NO	Indexed, Un-Tokenized	index-more	primaryType (mime-type)	?	?
subType	NO	Indexed, Un-Tokenized	index-more	subType (mime-type)	?	?
tld	YES	Un-Tokenized / NotStored (based on conf)	tld	Adds a **top level domain** field to the document.	?	?
subcollection	YES	Tokenized	subcollection	For Comprehensive description see src/java/org/apache/nutch/collection/**package.html**	?	?
urlmeta	NO	Indexed, Un-Tokenized	urlmeta	Adds any specified **url metadata tags** to the document in the index.	?	?

Jira Issues about indexing and IndexingFilterPlugins are

- index-extra plugin
- index-static plugin
- index-replace plugin

The index plugins to include are :

- index-(anchor | basic | more | static | replace) | tld | subcollection | creativecommons | language-identifier | urlmeta