WCA Crawler Plug-ins

Generic Crawler Plug-in

Highlights

Filtering Capability

"Data Source" Facet

"Directory" Facet

Custom Security Tokens

To extend the functionality of the *IBM Watson Content Analytics Crawlers*
IBM offers a set of special crawler plug-ins.

Generic Crawler Plug-in

The Generic Crawler Plug-in can be used to filter out documents from crawling based on URL patterns, allowing to configure e.g. crawling of a whole file system share with just filtering out documents inside distinct folders like "HR".

The module also provides a configurable way to create a "Data source" facet that contains a unique value for each crawled data source. This allows you to create a facet with values e.g. for "Filesystem", "Notes", "Intranet" and so on.

For directory based data sources like a filesystem the Generic Crawler Plug-in can generate data to build up a hierarchical facet resembling the directory structure of the crawled source.

For all data sources the Generic Crawler Plug-in can query additional access limitations from a JDBC database and add these as custom security tokens to the crawled documents, thus allowing to create an additional layer of security or to create custom security for data sources that do not offer secured document access.

Technical Information

Supported Systems

The current version of the service offering supports all crawlers included in *IBM Watson Content Analytics / IBM Content Analytics / IBM OmniFind* products, including custom crawlers build using the custom crawler framework.

Using an own crawler plug-in

Because only one plug-in per crawler is allowed, no other crawler plug-ins can be used natively. To extend the functionality of the Generic Crawler Plug-in with custom unctionality an own plug-in can be created that runs as a “child” of the Generic Crawler Plug-in module. Any existing crawler plug-ins have to be redesigned to run as such a "child" of the Generic Crawler Plug-in module.

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de

IBM Homepage is reachabel below:
ibm.com