

# Aspider Introduction

The Aspider Web Crawler connector will crawl content from any given website.

Aspider is based on the [Heritrix HTML Parser](#) for links discovery, but relies on the [Aspire 3 Connector Framework](#) to handle connections and distributed crawls. See [The Making of Aspider](#) for more information.

Aspider is highly configurable and behaves better for intranet crawls in comparison to the [Heritrix Crawler](#).

## On this page

- [Features](#)
- [Content Retrieved](#)
- [Limitations](#)

## Features

Some of the features of the Aspider Web Crawler connector include:

- HTTP Authentication
  - Basic/Digest
  - NTLM
  - Negotiate/Kerberos
  - HTML forms (cookie-based)
  - Connection throttling
- Incremental crawl
  - Ignore/Respect robots.txt and robots meta tags
- Heritrix HTML parser for link extraction
  - Connection proxy
- Configurable User agent
  - Max Crawl depth
- Distributed crawling
  - Include/Exclude patterns
- HTTPS crawling

## Content Retrieved

The Aspider Web Crawler connector retrieves several types of documents. Listed below are some examples of documents retrieved by this crawler.

- HTML pages
  - html, aspx, php, etc.
- Scripts and stylesheets
  - js, css, etc.
- Images
  - jpg, gif, png, etc.



This crawler will retrieve any document found linked in the HTML Markup as links (such as PDFs, MS Word, MS PowerPoint, etc).

## Limitations

Due to the design implementation, Aspider Web Crawler has the following limitations:

- Dynamic generated markup
  - Any markup generated by the browser by executing a site's javascript will NOT be detected by the crawler, so dynamic links will not be discovered.

Anything we should add? Please [let us know](#).