File System Extractor

Version 2.0

A great deal if not the majority of company information is still stored in files sitting on hard drives on PCs and network drives. This is fine so long as one knows where the data is, but as we all know it can become difficult to even locate files on our own systems. As hard drive space increases logarithmically and prices continue to sink and data continues to flood in, dealing with all of this information is outstripping any manual structuring or ordering that can be done. To combat this trend, the File System Extractor was developed.

The File System Extractor can crawl file systems quickly and easily. It is easy to setup and administrate with a unified graphical user interface even across multiple paths, drives, and systems. It handles all of the most commonly used file types and many more. It can then deliver the files in varied text formats (like XML) to fit the needs of CMSs, Search Engines, or other Content Platforms.

Additionally, the File System Extractor deals with all the details around handling the data that manual intervention is no longer needed. It polls the file systems where the data is to provide updates on changes, deletions, and additions of files. It extracts a wide breadth of valuable meta-data from the files properties. The File System Extractor does all of this to bring you and your organization's Intellectual Property back to your fingertips.

Supported File Systems

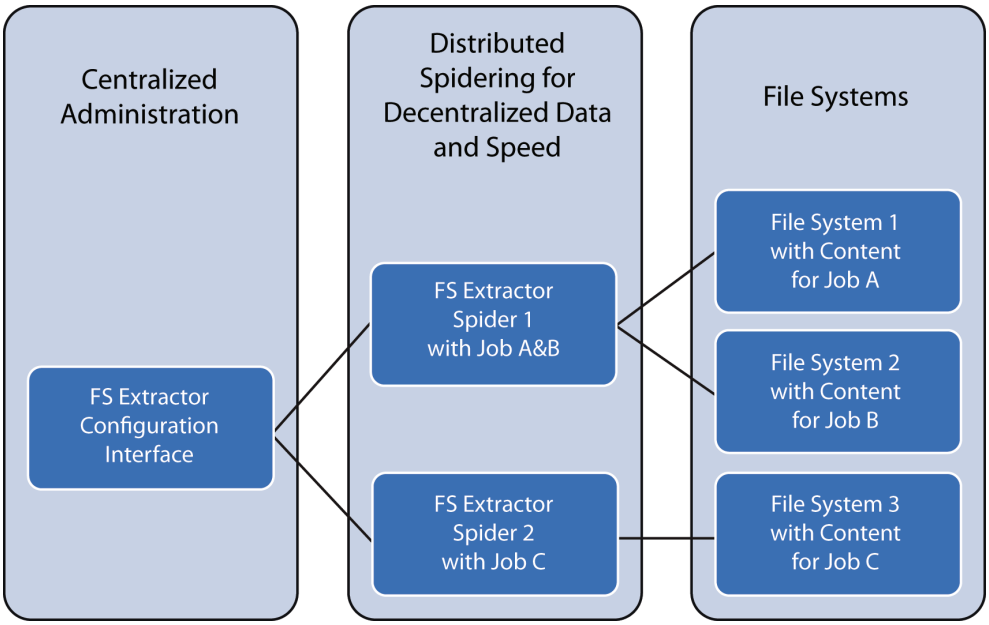
The File System Extractor uses SMB to connect to the various file systems. This makes it virtually Operating System independent as Windows, Unix, Linux, Mac, and almost all other Operating Systems support SMB connections.

Security

The native security of the files can also be used in the search engine or content platform using the text for of the file later. This is done through extracting the Access Control Lists (ACLs) of each file, and storing it in a specials fields which may be compared against later that only the files users should have access to may be seen.

Architecture

The architecture has been elegantly designed that it has a single point of administration while allowing multiple instances of the spider component to be run on several different machines. This allows scaling for large implementations even across multiple geographic locations while maintaining accuracy and ease of administration.



Setup and Administration

An often overlooked aspect of such a system is the effort involved in setting it up and administering it. The development team at 30 Digits have seen how complex and manual most of the solutions on the market are and have thus taken great effort to create a method for extracting content that is easy and intuitive while remaining flexible and powerful. Some of the main ways this has been achieved are by providing the following:

- Graphical User Interface (GUI) available via web browser for all of the configuration and administration options instead of textual configuration files which are often nearly impossible to decipher and lack documentation of how to setup and no central way to administrate them.
- The system uses AJAX extensively to load lists in the background to help fill out information.
- Simple field property modification like changing the title name of a field or even more complex operations like switching field variables from names to numbers for search engine optimization.

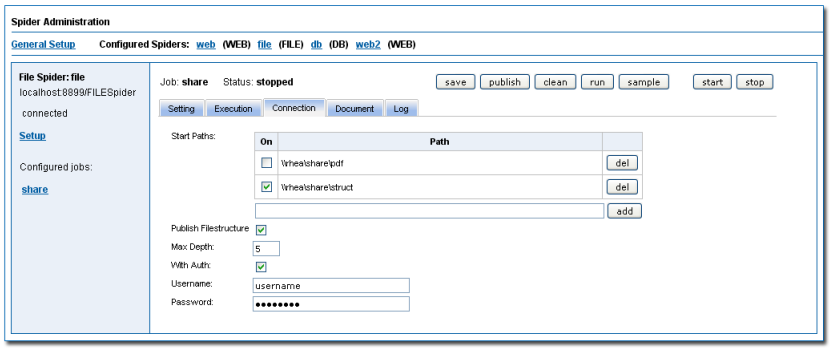
- Advanced features are available when an administrator wants to go deeper but not necessary for beginners or basic setup.

- Regular polling times for checking new data can be set very granularly and flexibly (i.e. every Monday, Wednesday, and Friday at 0, 6, 12, and 18 hours).

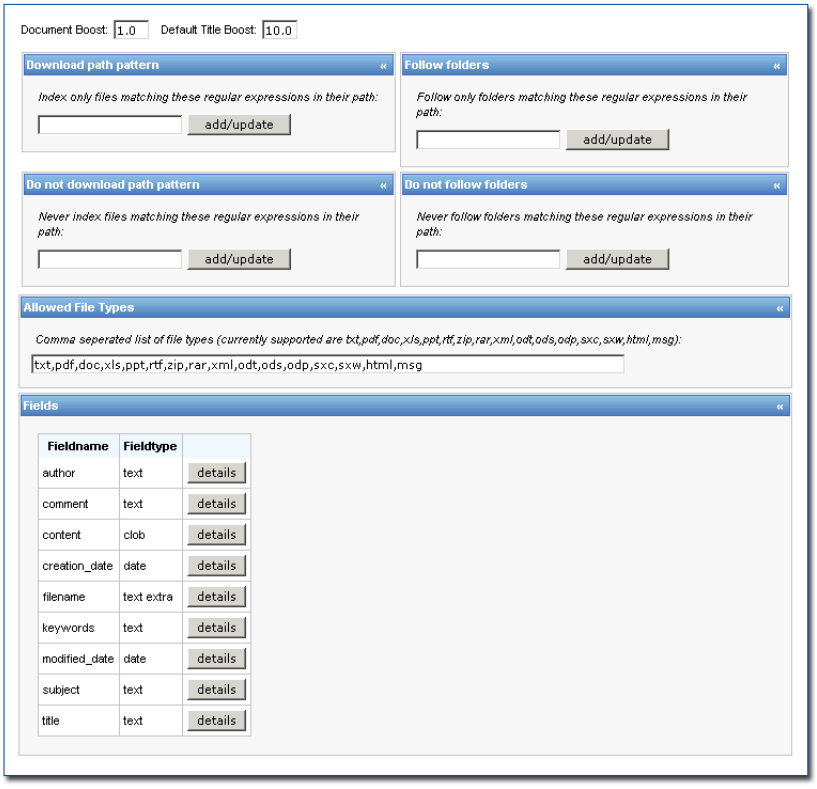
- When information is changed or deleted, the system can be set to either replace or delete the information in standard mode or keep time stamped versions of the document in the versioning mode.

- Load of spider on the sources can also be configured for faster spidering or less intensity on the server

- The File System Extractor can login to network drives requiring authentication by running the spider under the appropriate user or providing the login information in the configuration.



- Sampling of spider output before jobs are run. This can save tremendous amounts of time as different configurations can be tried before committing jobs that could have long run times. It also allows the capturing of errors at the beginning instead of waiting to analyze a log file after hours of spidering.



File types and Meta-Data extraction

Often the data around a file is also valuable. This can be the date and time it was saved or modified, the file name, the directory name it is saved in or other specific file properties like author, subject, title, and key words. This information is most often referred to as meta-data. The File System Extractor has the ability to extract this data and store them in fields associated with the main text of a document.

The File System Extractor also handles a large assortment of file types. Here is a list of some of the major ones: Microsoft Word, Excel, PowerPoint, PDF, HTML, TXT, RTF, MSG, XML, and ZIP (which are handled recursively to extract out the relevant data from each file within the ZIP). New file types are constantly being added to the import process to assure all kinds of files can be processed.

Contact

For more information or to schedule a demo, contact us at one of the following:

Tel: +49 89 45 23 89 66
Fax: +49 89 45 23 89 70
contact@30digits.com

30 Digits GmbH
Agnes-Pockels-Bogen 1
80992 München
Germany