Data sheet

HP WorkSite OCR Module

Find hidden information across the enterprise

Key benefits

- Boosts efficiency by finding hidden content including email attachments
- Seamless server side integration with HP WorkSite, and no desktop footprint
- Powered by IDOL with support for over 1000 file types and over 120 languages

Index the full text of all documents

In today's enterprise environment, your most critical information is often the 10% that has not been indexed, which means it lives outside the applications you use to effectively manage content. For instance, you may have thousands of valuable documents such as contracts, signed agreements, court documents, and other scanned content that are not full-text searchable because they were created by processes that do not include character recognition capabilities.

A few sources of image-only scanned documents:

- Ad hoc desktop scanners that do not have the ability to generate text for indexing
- Third parties who provide non-OCR-scanned content as email attachments
- Desktop OCR processes that lack enterprise throughput, failover, and error handling
- Internet research downloads or imports

Mitigate the risk of hidden information

Amassing image-only documents can create an unquantifiable risk for your organization. It can also mean that many critical documents, which contain a rich source of business information, are being underutilized. Since image-only information is visible just via navigation and metadata searches in HP WorkSite, these documents are not returned in full text search result sets because their content has not been indexed by HP IDOL.

Powered by HP IDOL, the WorkSite Optical Character Recognition (OCR) Module extracts the full text of these documents into the IDOL index collection, allowing you to search on the full contents of the document. This enables your organization to fully leverage the benefits of this content, moving it from being previously "hidden" to completely accessible.

Low cost of ownership

The plug-in nature of the module allows you to leverage existing IDOL infrastructure, eliminating the need for workstations or software on the desktop. Zero desktop footprint also means less overall maintenance for IT.

Powerful server-side processing

The WorkSite OCR Module is installed as a back-end service and performs two important functions:

- **Back file OCR:** The OCR module identifies image-only WorkSite documents, generates OCR text, and indexes the content.
- **OCR all incoming documents:** As part of the indexing process, the OCR module continuously extracts text from new and revised WorkSite documents.

No more hidden documents

Extracting content from the image files, including email attachments, supports smart business decision-making by providing the correct users with search access to this critical content. Once the documents are fully indexed, this important enterprise knowledge can be found, regardless of how they are searched within WorkSite.

Fast and powerful seamless automation

Since the WorkSite OCR Module is an IDOL plug-in that uses the existing IDOL indexing process, no middleware processes are involved, resulting in fast OCR processing. Documents can be made available for searching within minutes.

Image files added to, and any files already present in WorkSite, automatically become searchable as part of the normal IDOL indexing process, without any additional input or work from the end user or IT staff.

Accuracy across formats and languages

IDOL's ability to understand content in over 150 languages gives the HP WorkSite OCR Module the power to extract information from practically any document, regardless of its origin or language, with an unparalleled level of accuracy.

OCR is performed on all graphic files and documents (.pdf, .tiff, .jpg or .gif) regardless of size—a document can be one page or a collated set. Also, the OCR process is performed in place on the server side, so document integrity is always maintained.

Learn more at

http://www.autonomy.com/products/worksite

About HP Autonomy

HP Autonomy is a global leader in software that processes unstructured human information, including social media, email, video, audio, text, web pages, and more. Using HP Autonomy's information management and analytics technologies, organizations can extract meaning in real time from data in virtually any format or language, including structured data. A range of purpose-built market offerings help organizations drive greater value through information analytics, unified information access, archiving, eDiscovery, enterprise content management, data protection, and marketing optimization.

Please visit **http://autonomy.com** to learn more.

Sign up for updates
hp.com/go/getupdated

Share with colleagues

20141027_LC_DS_HP_WorkSite_OCR_Module