Business white paper

# Turn scans into editable or searchable text

**With built-in OCR Technology by I.R.I.S©**

**HP LaserJet Enterprise flow MFP M525c**
**HP LaserJet Enterprise color flow MFP M575 new flow MFPs**
**HP Digital Sender Flow 8500 fn1 Document Capture Workstation**
**HP LaserJet Enterprise flow MFP M830z**
**HP Color LaserJet Enterprise flow MFP M880z/z+**

# Table of contents

# Do more with data, using built-in OCR and multiple scan-to options

Empower work teams and help boost productivity with the feature-rich, HP LaserJet Enterprise flow MFPs. With a 100-sheet automatic document feeder, two-sided, single pass scanning, ultrasonic double-feed detection and onboard optical character recognition (OCR) processing,[1] these multifunction printers are your workflow performance expressway.[2]

The HP Digital Sender Flow 8500 fn1 also include OCR functionality. This document generally applies to these products as well.

This white paper will help you understand the capabilities and limitations of both OCR processing at the device and OCR technology itself. This information can help guide you as you implement OCR capabilities into your workflows.

## Just the basics

**What is embedded Scan to OCR?**
Optical character recognition (OCR) allows you to convert scanned images to common file types with editable and searchable text. OCR capabilities also enhance the indexing and retrieval of documents. When using embedded OCR, OCR processing occurs within the device itself.[1]

**How does it work?**
OCR software analyzes and converts the patterns found in a digital image of a page of text into text characters and saves them in a format that computers can search or index, such as searchable PDF, RTF (rich text format), Unicode, or ASCII text.

**How to obtain best results?**
Because the software is working hard to recognize the shapes of characters, it is important that the copy being scanned is as clean and crisp (high contrast) as possible. There are several additional factors that contribute to ultimate success in reading accuracy, which we will discuss in more depth later (see “Best practices: OCR tips for success” on page 11).

[1] Embedded OCR on the HP LaserJet Enterprise flow MFPs is intended for occasional use. For high-volume OCR usage, consider optional server-based Digital Sending Software (DSS).
[2] Other models in the MFP 525, color MFP M575 series, can also scan to OCR but require optional software, such as HP Digital Sending Software (DSS). HP Digital Sending Software, MFP M830, and color MFP M880 series are optional and must be purchased separately. See page 13 for more information.

## OCR Technology by I.R.I.S.

These MFPs use embedded OCR technology from I.R.I.S., a long-standing developer of OCR. With OCR, you can scan paper documents and turn them into fully editable text documents and text-searchable PDF documents. OCR is a key feature of more advanced workflows, and helps with document indexing and searching when sending to a variety of destinations, including network folders, email, SharePoint®, HP Flow CM Professional, or a USB flash (thumb) drive.

**Supported output file formats**
The MFPs support several file formats for outputting OCR, including:

- Searchable PDF (OCR) creates a PDF file with text you can search and select, while preserving the appearance of the scanned document.
- PDF/A (OCR) is a type of PDF designed for long-term archival of electronic documents. All formatting information in the document is self-contained.
- RTF (OCR) creates a rich text format (RTF) file. RTF is an alternative text format that can be opened by most word processing programs (Microsoft® Word-compatible). Some of the formatting of the original will be saved using this option.
- Text (OCR) creates an ASCII text (TXT) file that can be opened in any word processing program. ASCII provides little support for the expanded alphabet used by many non-English languages. The formatting of the original is not saved with this option.
- Unicode text (OCR) is an industry standard used to consistently represent text in any language. Languages that use non-Roman characters must use Unicode for TXT files. The formatting of the original is not saved with this option.
- CSV (OCR) uses the comma separated value (CSV) format. This type of file is recommended when reading spreadsheet documents and can be opened by most word processing spreadsheets or database programs.
- HTML (OCR) creates a hypertext markup language (HTML) file. HTML is used to display files on websites. (If the original document contains images, the resulting output will be stored in a zip file.)

**Supported OCR languages**
The built-in OCR software includes the ability to read the following 27 languages: Catalan, Chinese (simplified), Chinese (traditional), Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hungarian, Indonesian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Romanian, Russian, Slovakia, Slovenian, Spanish, Swedish, and Turkish.

You can change the OCR language from the control panel (instructions begin on page 10).

**How do I scan to OCR?**
There are a few ways to scan documents for OCR use: you can do everything from the color touchscreen control panel (see instructions beginning on page 5). You can also use a custom HP Quick Set to launch an OCR workflow for the settings you want every time. With Quick Sets, users can find what they need right away, without standing at the device control panel searching for the appropriate settings. See page 6 for Quick Set instructions.

## Scan to OCR from the control panel

To scan from the control panel without using custom Quick Sets:

**Changing defaults from the control panel**

You may want to change the default scan settings from the MFP's control panel. You can do so from the **Administration** menu, **Scan/Digital Send Settings**.


1. Place the document(s) to be scanned in the automatic document feeder (ADF), or on the scanner glass, and select the location where you want to save the scanned file (for this example, we will save to USB). Type a filename for your saved file and press **File Type**. A drop-down list appears.


2. Scroll down the list and select the OCR **file type** that best suits your needs.


3. If you want to customize scan settings, press **More Options** to specify a wide variety of settings, such as cropping, blank page suppression, quality, and much more.


4. Press the green **Save to**... button The MFP scans and processes the document. (You may opt to **Preview** the scan first. See "Using image preview to proof and optimize scans," page 11).


5. The control panel informs you of the job status along the way. When the scan is successfully completed, a Status message reads **Success** (or it will inform you of any problems).

### Finding your device's IP address

**MFPs:** To display the IP address on the Control Panel, touch the connect info icon. Digital sender: Touch the Administration button, then touch Reports.


### Note:

The administrator can disable the Info button on the Control Panel. If the button is not present, you can find the IP address in the Administration menu (Administration > Reports > Configuration/ Status Pages > How to Connect Page). If that is also unavailable, contact your administrator for the IP address.

### Certificate error?

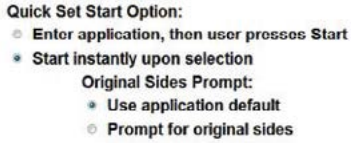
If you see a security certificate error, click **"Continue to this website (not recommended)."** (Don't worry, it's perfectly safe!)


### Quick Set button location and option:

- **Button Location.** This dropdown lets you select whether your new Quick Set will appear as a new item on the Home Screen or under the Quick Sets button.


Quick Set Start Options:

- **Enter application, then user presses Start** causes the device to pause and wait for your input on the scan operation. **Start instantly upon selection** scans whatever is on the ADF or on the glass as soon as you press the Quick Set button (unless you set the Prompt for original sides option).

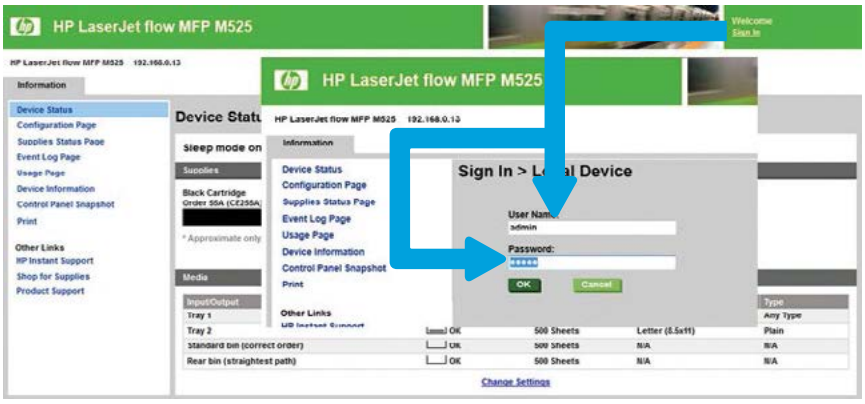



## Set up a custom Scan to OCR HP Quick Set

Custom HP Quick Sets can be created using the device's embedded web server (EWS). For this example we will create a Save to USB Quick Set, but you may want to send your output to a network folder, an email recipient or another destination, such as SharePoint. (Note that any settings in the Quick Set can be overridden at the control panel or modified later in the EWS to optimize OCR results if needed.)

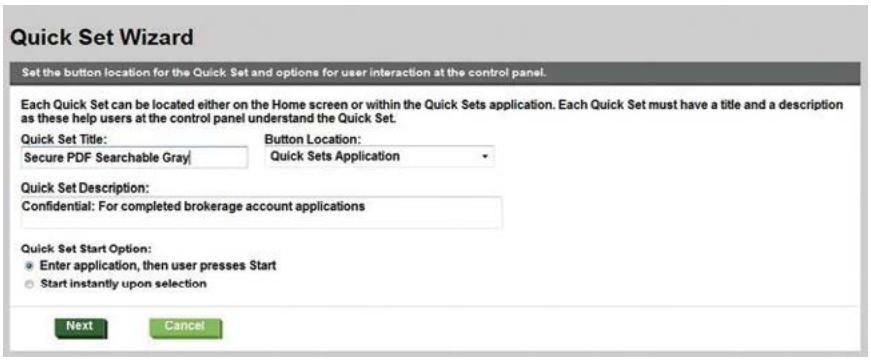
1. Using a browser, such as Microsoft Internet Explorer, Open the EWS by entering the device's IP address into the address bar (see note at left).

2. The home page opens. Click **Sign in** at the upper right corner and enter your administrator credentials (admin account and password).




3. Click the **Scan/Digital Send** tab (A). From the left navigation menu, select the desired destination for your scanned output files (B). If it is not already selected, check the selection box, **Enable Save to**... (C), or you can opt to wait to enable, then click **Add** (D).

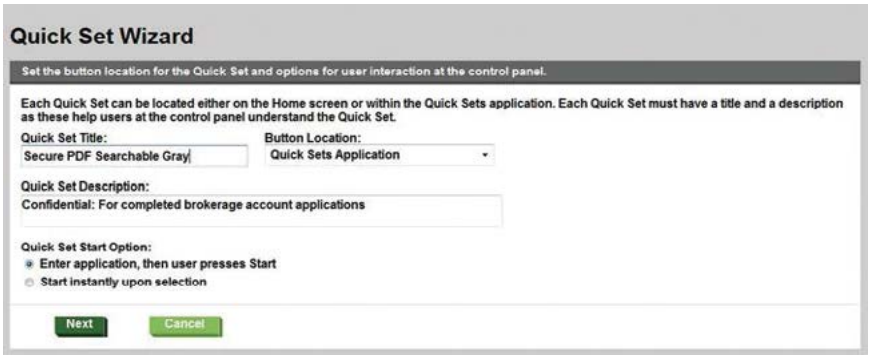

4. A **Quick Set Wizard** dialog box opens. Supply a title and description for the OCR Quick Set, select the **Button Location** for the Quick Set and the **Start Option** for user interaction at the control panel, and then click **Next**.



**Folder creation on the fly**

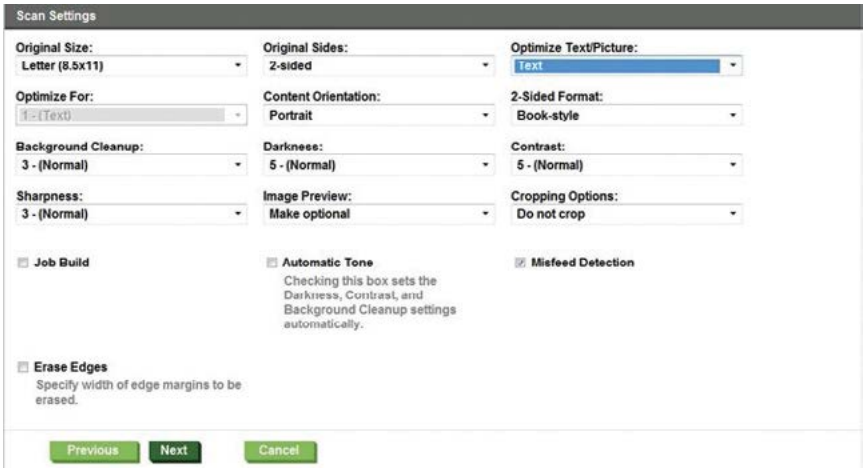
If you specify a folder or path that does not yet exist, the first time you use the Quick Set, the MFP will ask if you want to automatically create it.


5. The next screen provides options for determining where the files will be saved. For this example, we are specifying a folder named “Brokerage Apps” under the folder “Confidential.” (In the Quick Set wizards for sending to other locations, such as to a network folder, there would be a different set of options pertaining to folder settings.)




6. The next screen allows you to set a notification setting which will send an email notifying the user either upon job completion or only if the job fails.


7. Select any additional **Scan Settings** desired and then click **Next**.



**More about file settings**

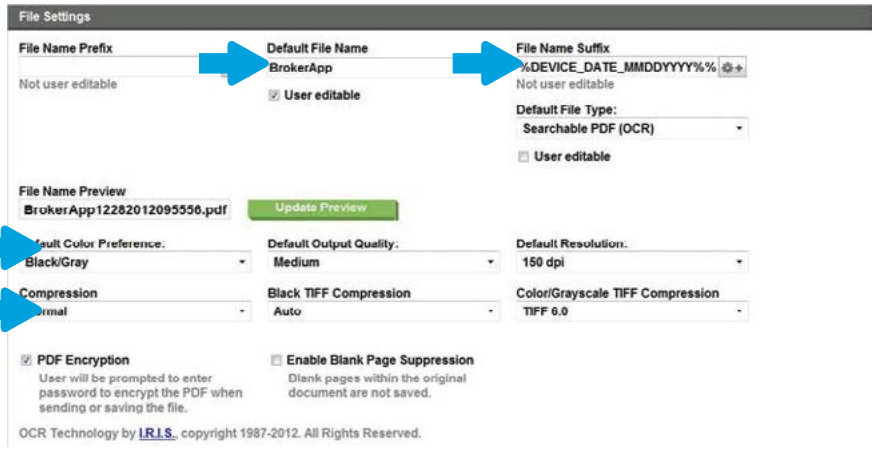
- **File Name Prefix/Suffix** options help organize files by adding identifying information to the beginning or end of the filename. Use **File Name Preview** to check the results.
- **Default Resolution** determines the **output** resolution of your scan in number of dots per inch (dpi). You can override this default setting from the control panel at the time of the scan. Generally speaking, the higher this setting, the larger the output file.

**NOTE:**

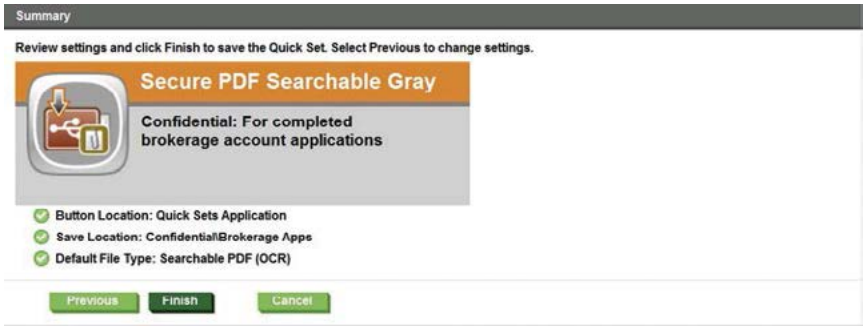
For OCR scanning, the MFPs will always **acquire** the image at 300 dpi, regardless of this **output** setting. So increasing this setting will not improve OCR performance. However, higher resolution settings may improve graphics quality.

- **PDF Encryption**—use this option if you need a secure, encrypted PDF file. You will be prompted to create a password before sending, that must be used to open the file.
- **Enable Blank Page Suppression**—this is a useful option if you scan documents that have blank pages. PDF files especially will be much smaller as a result of discarding blank pages.

8. Select any **File Settings** desired. For this example, we set the **Default File Name** to BrokerApp with a **File Name Suffix** containing the date and time. We set the **Default File Type** to **Searchable PDF (OCR)** and the **Default Color Preference** to Black/Gray because we do not need color scans. We left the default resolution setting of 150 dpi.




9. Review the **Summary** dialog, and if all looks well, click **Finish**.




10. A confirmation dialog appears confirming the successful creation of the Quick Step, or any additional action required. Click **Apply**, and if you checked the **Enable Save to**... box in step 3 on page 6, your new Quick Set will now appear on the device's control panel, ready for use.

## Scan OCR using a custom Quick Set

Once you have defined a custom OCR Quick Set and applied the changes from the EWS, it will appear on the device's control panel.

1. If you selected a button location of "Quick Sets Application" in step 4 on page 6, press the **Quick Sets** button to find your Quick Set. (If you selected "Home Screen" your new Quick Set will appear on the home screen instead.)


2. Place the document(s) you want to scan in the ADF input tray or on the flatbed scanner glass and press your Quick Set. Depending on the settings you specified, your scan job may process immediately, or, if you selected the Quick Set start option **Enter application then user presses start** (step 4 on page 6) you will see a screen like the one shown below. When ready, Press the green **Save to**... button. The MFP scans and processes the document. (You may opt to Preview the scan first. See "Using image preview to proof and optimize scans," page 11).


3. The control panel informs you of the job status along the way. When the scan is successfully completed, a **Status** message reads **Success** (or it will inform you of any problems).

**More about setting the OCR language**

At the time of this writing, the only method for specifying the OCR language is from the device's control panel.

Future firmware releases may add the ability to set OCR language in a Quick Set using the Embedded Web Server (EWS).

Note that it is also possible to change the default OCR language for **all scans** to a specified destination, as follows.

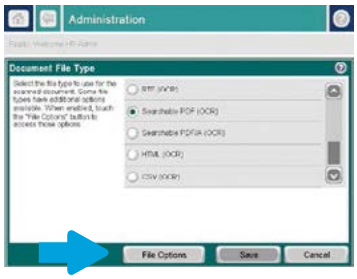
**To change default OCR language settings:**

1. From the control panel, select the **Administration** menu button (may prompt for access code).

2. Select Scan/Digital Send Settings.

3. Select the scan destination where you want to change the default language.


4. Select **Default Job Options** then **Document File Type**, then press **File Options** and select the language of your choice. All OCR scans will now default to read that language.




## How to specify the OCR language

If you selected Enter application then user presses start (see step 4 on page 6), you can easily specify the OCR language before scanning the document.

1. Place the document(s) to be scanned in the ADF (or on the scanner glass) and select the location where you want to save the scanned file (Network Folder, USB, SharePoint, etc.). Type a filename for your saved file and press **File Type** (even if the file type is already correct—doing so will allow you to change File Options). A drop-down list appears.


2. Press the green **Save to**... button. The device scans and processes the document. (You may opt to Preview the scan first. See "Using image preview to proof and optimize scans," on page 11).


3. When your scan is complete, you will have the option to **Retain settings for next job**. Check the box if you will be immediately scanning more documents using the same destination and scan settings.

# Best practices: OCR tips for success

**What can and cannot be OCR'd**
OCR technology has advanced significantly in recent years. However, successful OCR scanning depends on a number of factors, including the properties of the original document. For example, when the original is clean and crisp, and without contrasting watermarks or other distractions, the accuracy of the resulting text can be quite high. If you are uncertain of the suitability of source documents, you may want to conduct experimentation to judge trade-offs between OCR and keyboarding options. This should help you develop an OCR strategy that fits your document workflow needs.

**Following are some examples of scanned documents that should yield highly accurate results:**

- Clean and crisp text on white paper in a commonly used non-decorative typeface
- Documents without contrasting watermarks, colored backgrounds or other distractions

**OCR should not be attempted on certain materials, for example:**

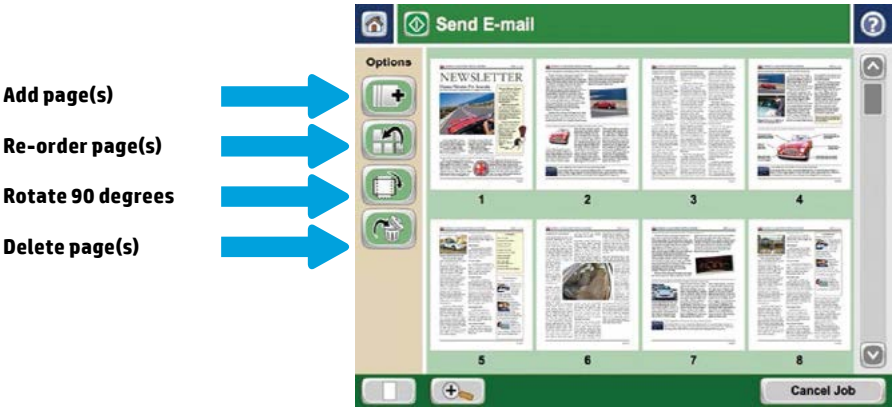
- Bar codes
- Handwritten text or stylized type that mimics handwriting
- Decorative or archaic typeface that is not in common use
- Poor-contrast documents, documents with faded print or printed on colored paper
- Documents that have been folded (text that has a crease running through it may cause misreads)
- Font sizes smaller than 8 points

**Suggested settings for best results**

- If some of your documents include blank pages (for example, a double-sided document with an odd number of pages resulting in a final blank page), turn on **Blank Page Suppression** to keep file sizes to a minimum. This is especially important when scanning PDF files.
- **Crop to content** can also help reduce the size and speed processing of PDF files.
- If you do not need color, set the scanner to black or **black/gray**. This can reduce processing time and file size.

**Using image preview to proof and optimize scans**
The control panel on these devices features image preview, which lets you minimize steps and reduce errors. You can view and edit your scan job right on the device's color touchscreen control panel before you send it to a network folder, fax, email, or USB device. You can easily get your scans to look just the way you want them. Instantly preview and modify scans before sending them into the workflow, eliminating the need to walk back and forth from a computer.

Scan multiple pages using the ADF and view them as thumbnail images, or zoom and pan for a detailed inspection. Before finalizing a document you can add a page to the scan, reorder pages, rotate a page or pages, or delete unwanted pages. You can apply changes to a single page, or multiple pages at once.



**Informal testing limitations**

Note that these example scan results are intended to provide a very generalized picture of relative file sizes and scan times for a typical sampling of document types. The data may help you in planning your workflow in terms of file storage requirements and time concerns but are not intended as a definitive guide.

Actual results will vary, depending on a number of factors, including the specific document being scanned, and the scanning device and firmware version used (see Test Scan Conditions, below). HP is continually developing firmware to add features and adjust tradeoffs in performance vs. file size and image quality.

# OCR processing time and file sizes

Using OCR while scanning requires more processing time than without OCR.

**Example scans**

The table below shows a comparison of processing time and file sizes for the various OCR file types as well as two non-OCR scans of sample documents. These test scans were performed on an HP LaserJet Enterprise flow MFP M525c. Network scanners, future firmware releases, and future flow MFPs will produce different results.

- **All text:** a five page document with no graphics or photos, using a 10 point Arial font (san-serif).
- **Text and photos:** five pages of text, bulleted text, a table, and a few medium size photos.
- **Text and graphics:** five pages of text with graphics, taken from the pages of this white paper.
- **Spreadsheet:** one page spreadsheet with 17 columns and 25 rows of data, in an 8 point font.

| | | | OCR file formats | | | | | Non-OCR | |
|---|---|---|---|---|---|---|---|---|---|
| **Document type** | **Searchable PDF** | **Searchable PDF/A** | **RTF** | **Text** | **Unicode text** | **HTML[3]** | **CSV** | **PDF** | **XPS** |
| **All text** (5 pgs, black) File size (KB) | 230 | 190 | 27 | 17 | 33 | 38 | n/a | 103 | 196 |
| **Processing time** (min:sec) | 1:59 | 2:05 | 1:27 | 1:21 | 1:18 | 1:29 | n/a | 0:11 | 0:12 |
| **Text and photos** (5 pgs, black) File size (KB) | 1016 | 1590 | 165 | 18 | 36 | 183 | n/a | 151 | 257 |
| **Processing time** (min:sec) | 2:41 | 2:59 | 2:06 | 1:28 | 1:28 | 2:00 | n/a | 0:16 | 0:16 |
| **Text and photos** (5 pgs, color) File size (KB) | 1144 | 1744 | 164 | 18 | 18 | 224 | n/a | 1515 | 1540 |
| **Processing time** (min:sec) | 2:57 | 3:07 | 2:26 | 1:52 | 1:46 | 2:21 | n/a | 0:13 | 0:16 |
| **Text and graphics** (5 pgs, black) File size (KB) | 556 | 887 | 355 | 12 | 23 | 404 | n/a | 713 | 769 |
| **Processing time** (min:sec) | 2:53 | 2:56 | 2:31 | 1:57 | 1:54 | 2:20 | n/a | 0:17 | 0:21 |
| **Text and graphics** (5 pgs, color) File size (KB) | 623 | 924 | 436 | 10 | 19 | 424 | n/a | 742 | 808 |
| **Processing time** (min:sec) | 3:25 | 3:39 | 2:53 | 2:13 | 2:09 | 2:19 | n/a | 0:17 | 0:21 |
| **Spreadsheet** (1 pg, black) File size (KB) | 115 | 59 | 60 | 4 | 4 | 206 | 9 | 25 | 37 |
| **Processing time** (min:sec) | 0:30 | 0:39 | 0:25 | 0:21 | 0:23 | 0:23 | 0:22 | 0:07 | 0:08 |

**Test scan conditions**

All scan tests were performed on an HP LaserJet Enterprise flow MFP M525c, with firmware version FutureSmart 2 SP1.11 (2201002 231113). Each of the scan tests represented in the table were performed one time only and are not averaged over multiple scans. Thus the information presented is not intended as comprehensive test data.

[3] HTML output creates a zipped folder. File size noted is the actual uncompressed html file size along with any graphics files.

For the text and text and graphics documents, we selected three typical desktop publishing documents. All these originals were five pages in length, double–sided, which left the sixth page blank. The spreadsheet was one page, single-sided. Scan settings used are shown below.

- For all scans: scan two sides; crop to content; blank page suppression; scan quality medium, scan resolution: 150 dpi.[4]
- For the "All text" documents: black only, optimized text.

## About optional HP Digital Sending Software

While OCR capabilities are built-in to the HP LaserJet Enterprise flow MFPs and HP Digital Sender Flow 8500 fn1, HP also offers optional HP Digital Sending Software,[5] recommended when scanning to OCR in high volume. HP DSS enhances digital sending across a wide range of HP LaserJet MFPs and DSS is only compatible w/ Milano / HP Digital Sender Flow 8500 fn1. HP DSS provides a common administrative utility and end-user interface across multiple devices and device types. This optional software runs as a service on a network server and allows devices to send jobs through the server. It is not necessary to install any software or drivers on individual users' computers.

HP DSS enables the following sending features:

- OCR allows you to convert scanned images to common file types with editable and searchable text. File types include all those listed earlier in this document (additional OCR languages include Arabic and Hebrew)
- Remote copy (sending to a printing device)
- Scan to shared folders, multiple folder destinations at one time and scan to Microsoft SharePoint
- Custom keys allow users to effortlessly send documents to workflow destinations
- Configuration utility for managing digital sending and workflow capabilities and manage multiple devices with configuration templates
- Central email routing
- Central address book management where users gain access to their Microsoft Exchange contacts, as well as their private address books, including secure email using the SSL protocol
- LDAP replication allows DSS to offload the LDAP directory activity by replicating relevant addressing information into the DSS address book

For more information about HP DSS, please visit hp.com/go/dss.

[4] All OCR scans are acquired for processing at 300 dpi, regardless of the output resolution setting. But the resulting file will be saved at whatever setting is specified.

[5] HP Digital Sending Software is optional and must be purchased separately.

**More resources**

**See various MFP simulations:**
**hp.com/sbso/product/mfp/demo/m575.html**
**hp.com/sbso/product/mfp/demo/m525.html**

**See a flow MFP introduction video:**
**http://www8.hp.com/h20621/video-gallery/us/en/products/scanners-and-fax/scanners/2682865856001/accelerate-productivity/video/**

**See an Enterprise network scanner video:**
**www8.hp.com/h20621/video-gallery/us/en/products/scanners-and-fax/scanners/1377742364001/hp-scanjet-document-capture/video**

**See a flow MFP marketing video:**
**youtube.com/watch?v=U0QuJOiOj1Y**

## In conclusion

OCR scanning provides numerous benefits, including the ability to edit text, and to index and search archived documents. Built-in OCR on the HP LaserJet Enterprise flow MFPs and digital sender is enhanced by the innovative scanning capabilities of these versatile multifunction devices, such as the 100-sheet automatic document feeder, two-sided, single pass scanning, ultrasonic double-feed detection, and advanced image processing.

HP Quick Sets let you launch document workflows and get it right—at the touch of a button. HP Quick Sets help by automating all of the steps of a complicated workflow, all on a large, full-color touchscreen that is a joy to use. Users can find what they need right away, without standing at the device control panel searching for the appropriate settings.

When integrating OCR scanning into your workflow, you should consider both the limitations of OCR technology and the performance parameters of the hardware. Using the information contained in this document will help you to optimize your OCR results.

**Learn more at**
**hp.com/go/flow**

**Sign up for updates**
**hp.com/go/getupdated**

Share with colleagues

Rate this document



4AA4-5159ENW, November 2013, Rev.2