

Getting Started with AutoPortfolio™

Plug-in for Adobe® Acrobat®

Starter Guide to Converting and Processing PDF Portfolios

Contents

What is AutoPortfolio?	2
Converting a PDF portfolio into a single, “flat” PDF document	2
Extracting all files from a PDF portfolio	2
Extracting PDF Portfolio metadata into Excel spreadsheet	2
Converting a PDF Portfolio into TIFF image and text format.....	2
De-duplicating PDF and load files	2
Converting a PDF Portfolio into a single PDF document.....	3
Extracting All Files from a PDF Portfolio	12
Extracting PDF Portfolio metadata into Microsoft Excel Spreadsheet	18
Exporting PDF Portfolio into TIFF and Text format.....	21

What is AutoPortfolio?

The AutoPortfolio is a plug-in (add-on) for Adobe® Acrobat® software. It is designed to perform the following operations:

Converting a PDF portfolio into a single, “flat” PDF document

- All portfolio items/documents are placed into a single PDF document.
- File attachments can be optionally converted to PDF and placed right after parent documents.
- MSG and ZIP files are automatically extracted and their contents processed.
- The ability to process entries only for a selected date or a person (as in the case of email).
- The output document is hierarchically bookmarked and is ready to be Bates stamped.
- Portfolio metadata is exported into a number of spreadsheet ready formats.

Extracting all files from PDF portfolio

- All items/documents are extracted as separated files into a single output folder.
- File attachments can be optionally converted to PDF and appended to parent documents.
- The ability to process entries only for a selected date or a person (as in the case of email).
- The contents of MSG and ZIP files are automatically processed.
- Portfolio metadata is exported into a number of spreadsheet ready formats.

Extracting PDF Portfolio metadata into an Excel spreadsheet

- All metadata information is extracted into an Excel spreadsheet (XML or text CSV formats). If processing email portfolio this includes all email fields such as “From”, “To”, “Subject” and etc.
- The ability to extract information only for a selected date or a person (as in the case of email).
- MD5 checksums are automatically computed for all files.

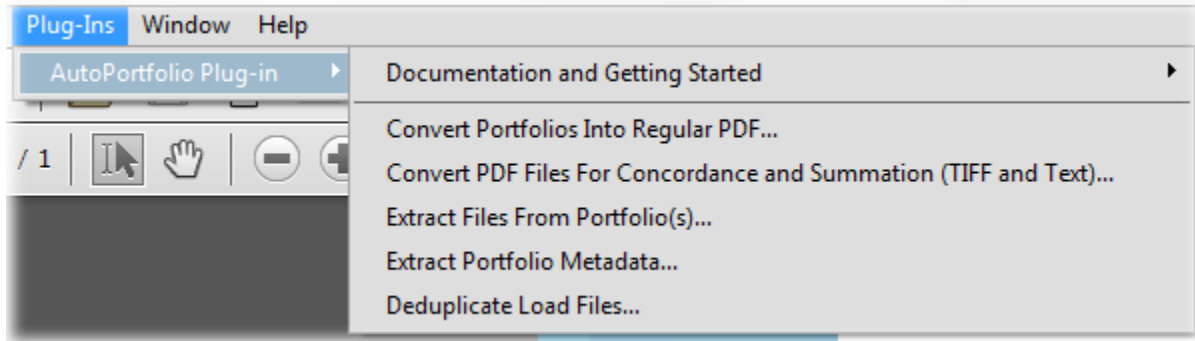
Converting a PDF Portfolio into TIFF image and text format

- Output is suitable for importing into litigation support systems such as Concordance and Summation.
- The ability to process entries only for a selected date or a person (as in the case of email).

De-duplicating PDF and load files

- Finding duplicate and near duplicate PDF documents.
- Specifically designed for comparing and de-duplicating emails.
- Significantly reduces the amount of documents that need to be printed or processed.

All operations are available from the main Acrobat menu via the “Plug-ins > AutoPortfolio” menu.

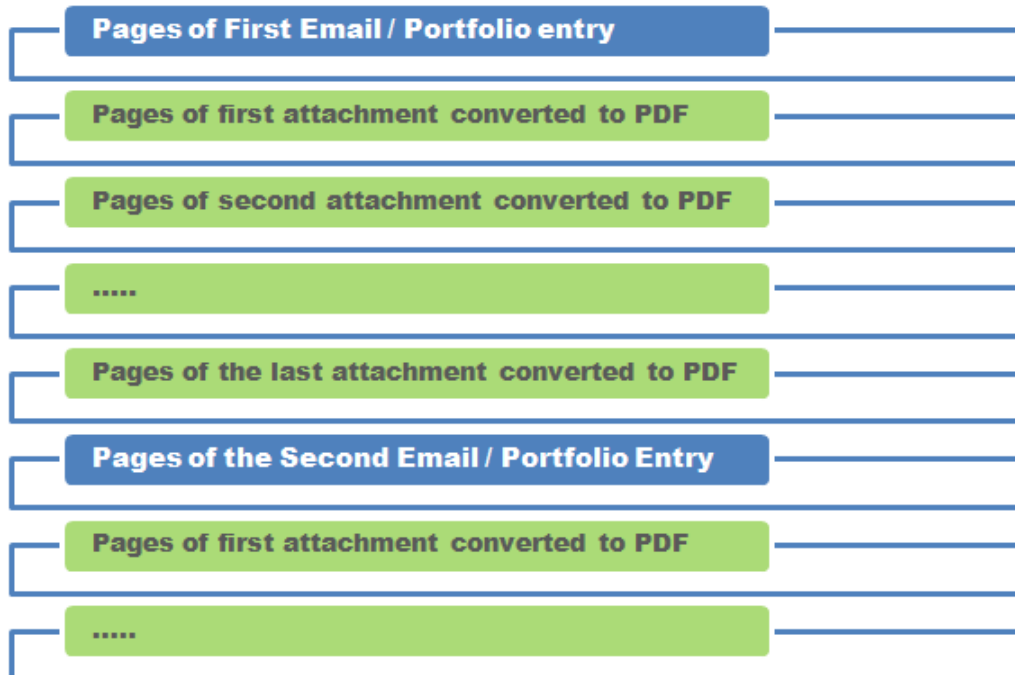


Do not open PDF Portfolio directly in Adobe Acrobat – use the plug-in menu to start AutoPortfolio. If a PDF Portfolio is open in Adobe Acrobat, then the “Plug-ins” menu is automatically disabled. Every operation provides a way to select one or more input PDF portfolios.

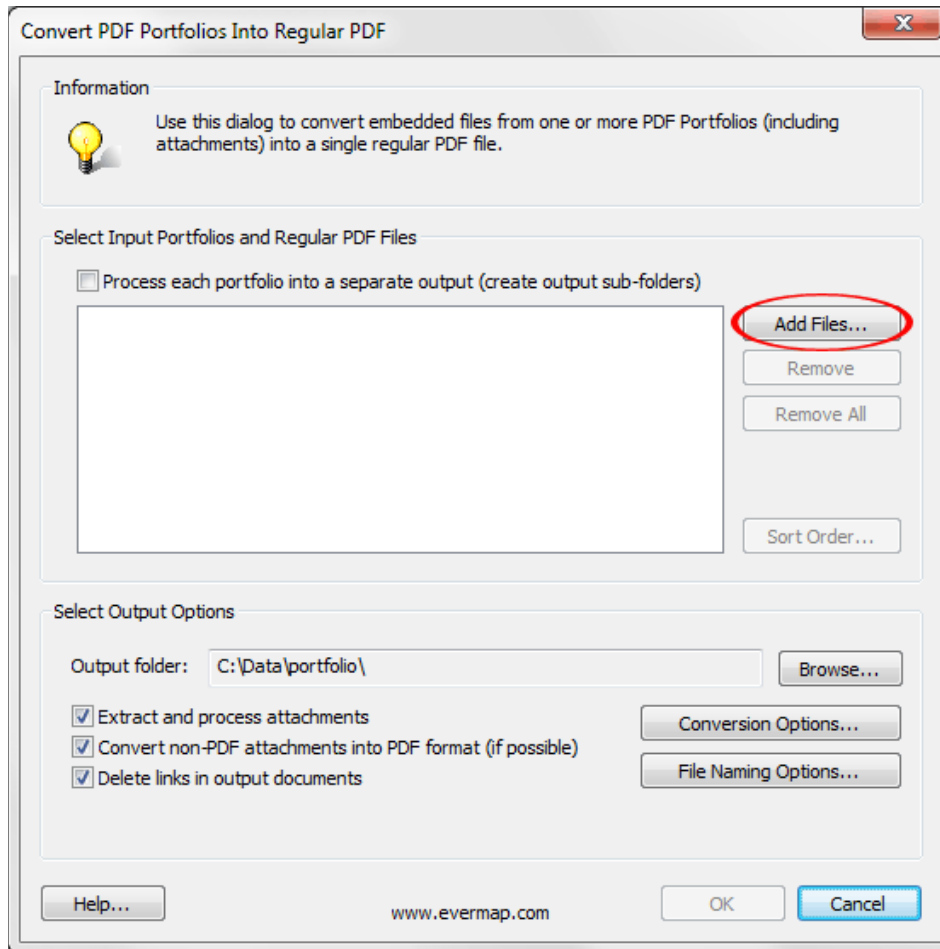
Converting a PDF Portfolio into a single PDF document

This operation takes one or more PDF Portfolios and converts them into a regular "flat" PDF document. All portfolio attachments can be optionally converted to PDF format and appended right after their parent document. Every document and attachment is hierarchically bookmarked in the output PDF document. Metadata and page index file is exported into a number of spreadsheet-ready formats. If an attachment cannot be converted into PDF format, a stub page is generated and inserted into output.

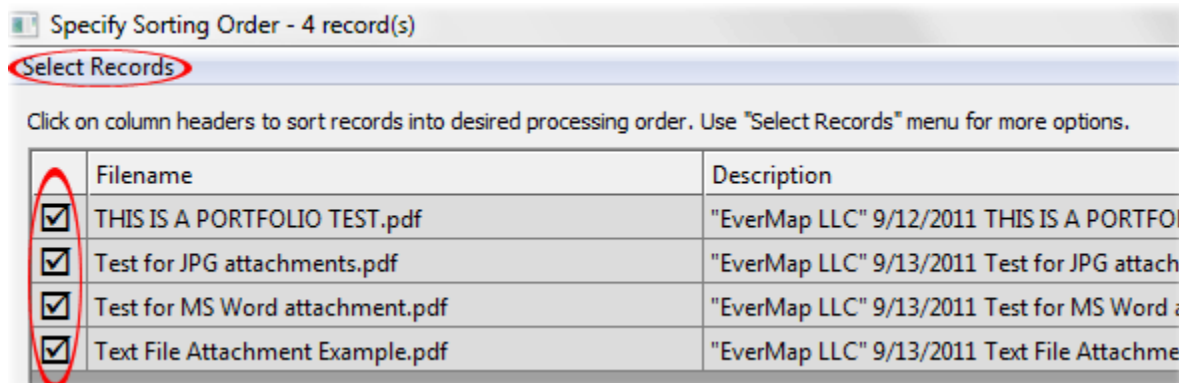
Here is a logical structure of the output PDF document for a sample PDF Portfolio that contains emails:



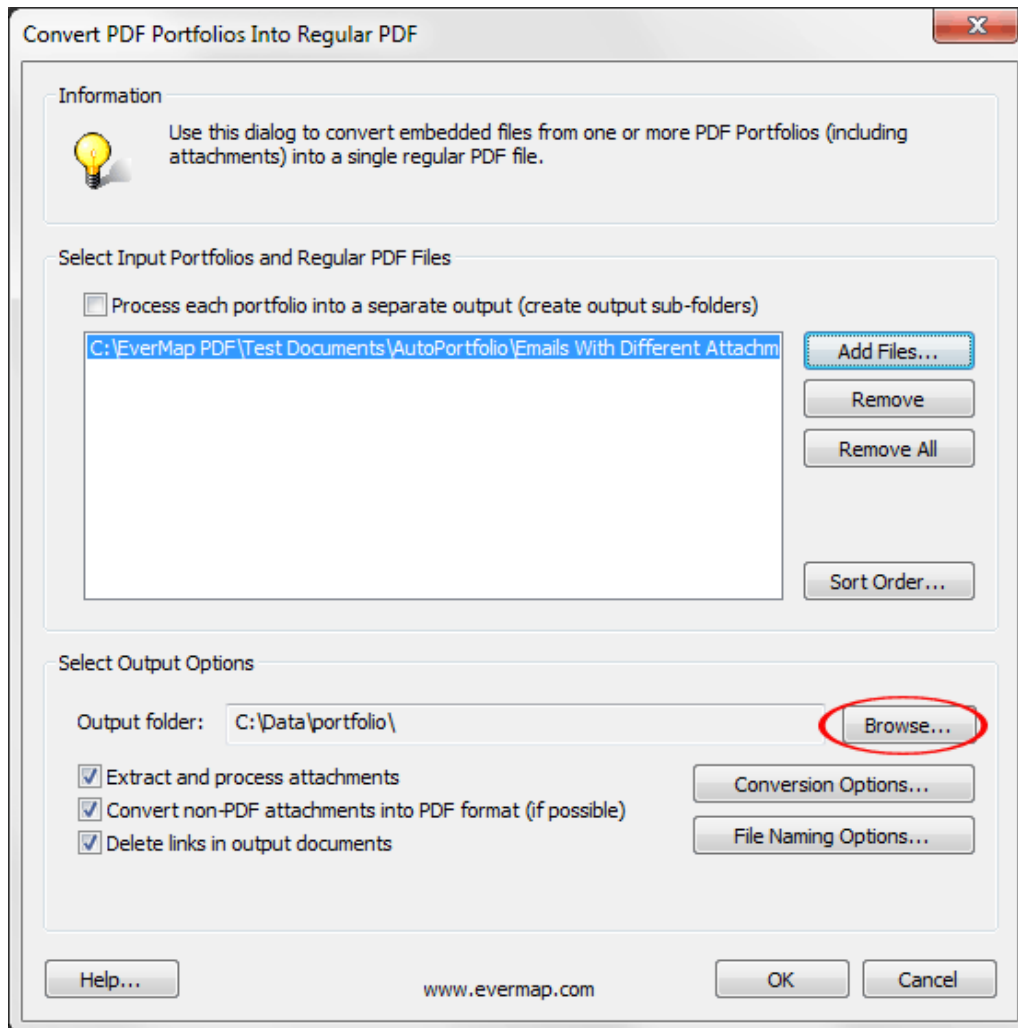
Select “Plug-ins > Convert Portfolios into Regular PDF” from the main Adobe Acrobat menu to open a conversion dialog (do not open a portfolio directly in Acrobat):



Press the “Add Files...” button to select the input PDF portfolio for processing; the “Specify Sort Order” dialog will appear on screen. This dialog allows for the selection of only a part of the PDF Portfolio for processing either by checking a box in front of every record or by performing a text search. If you want to select only a specific subset of the portfolio entries for processing, then use the “Select Records” menu to manipulate the current selection set.



If you want to process the entire portfolio, then simply press the “OK” button located in the lower-right corner of the screen. Select an output folder by pressing the “Browse...” button:



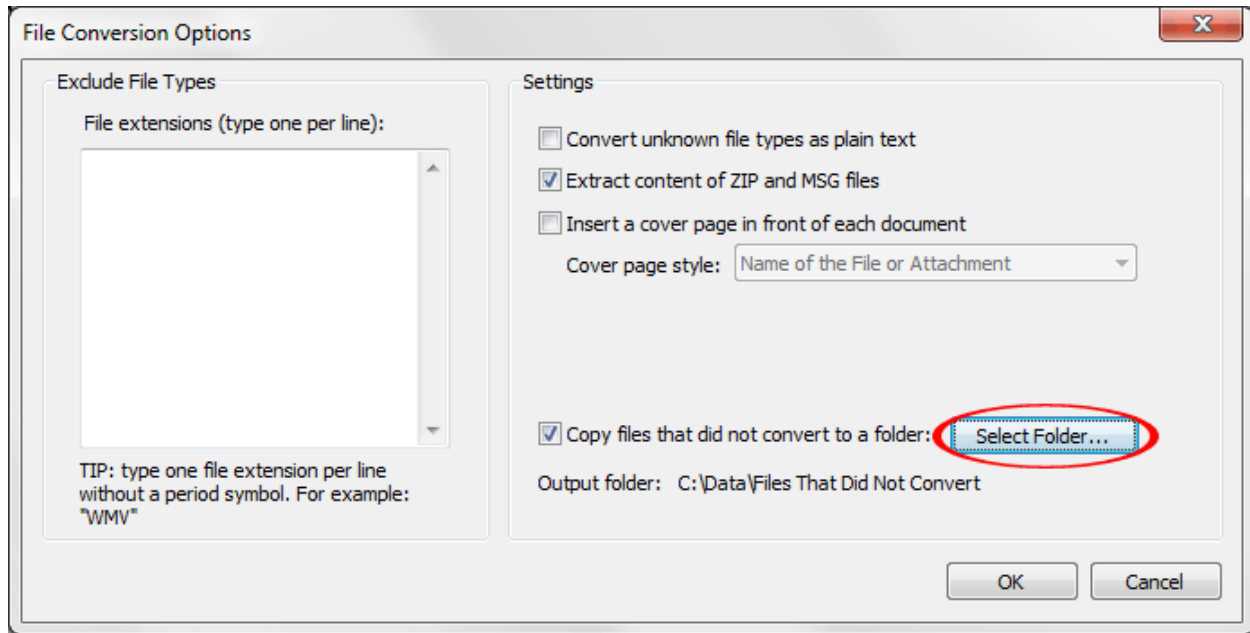
It is recommended that you select a short output path (like c:\Data\) located on the local hard drive for better performance.

Check the “Extract and process attachments” and “Convert non-PDF attachments into PDF format” options if you want to process file attachments for each portfolio entry. For example, if a PDF Portfolio was created by exporting emails from Microsoft Outlook, then these options will convert email attachments into PDF format and append them right after the parent email. If a certain attachment cannot be converted into PDF format (for example, a file is password-protected or is not in a file format that can be converted into PDF), then a stub-page will be inserted into the output file stating this fact. The stub page will show the attachment name and file size. Every attachment is hierarchically bookmarked for easy navigation and preservation of the original document structure.

Check the “Delete links in output documents” option if you want to completely remove the file links to file attachments in their native file formats as they are originally stored within the PDF Portfolio. For example, if an email contains a Microsoft Word attachment, then in the output file, the link to the MS

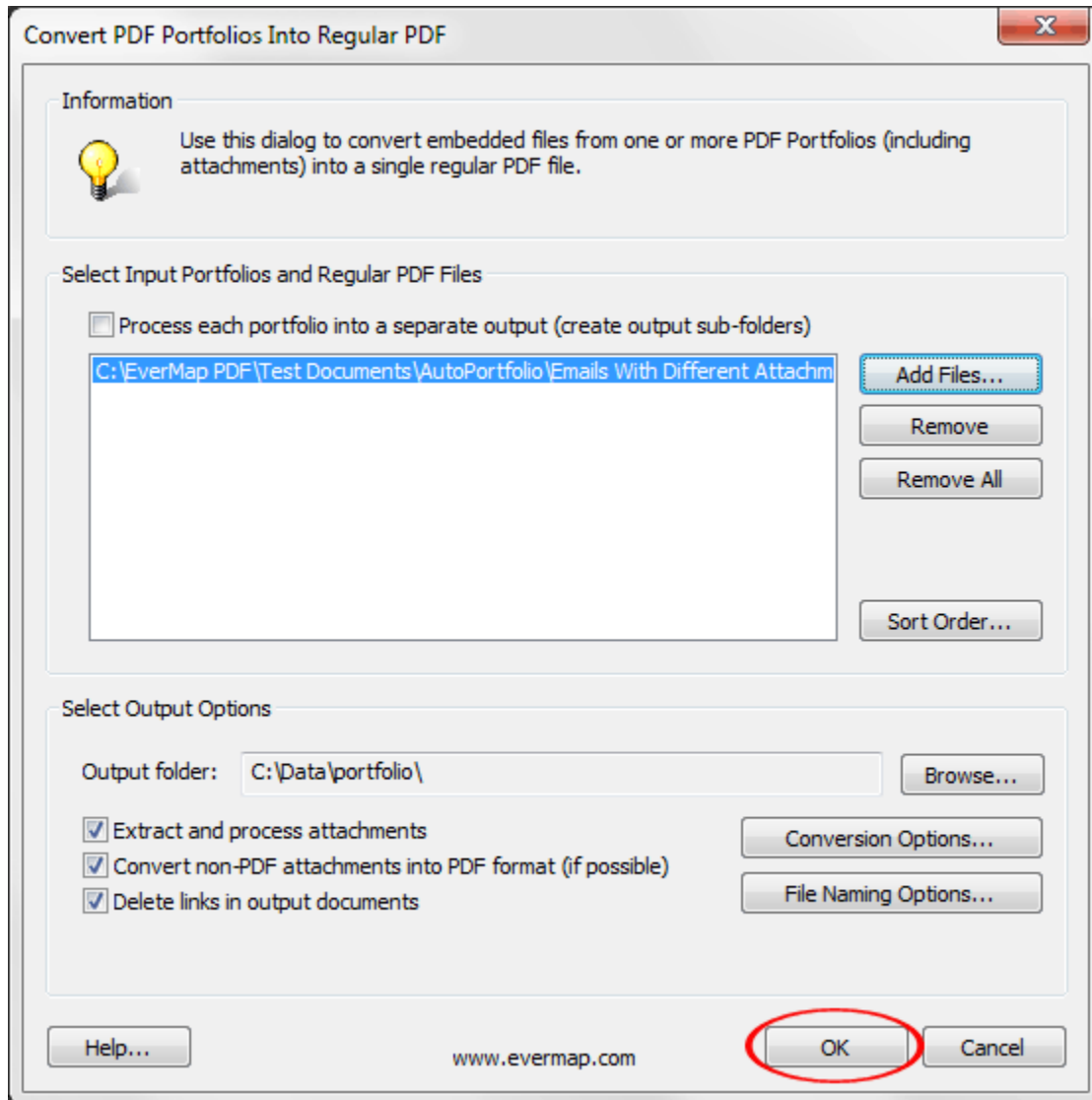
Word document will be removed. The attachment will only be available in its PDF representation (if the “Extract and process attachments” option is selected). The output file size may be significantly decreased if this option is used, since all attachments are not going to be stored twice (in the native and PDF formats).

Click the “Conversion Options...” button to open the “File Conversion Options” dialog.

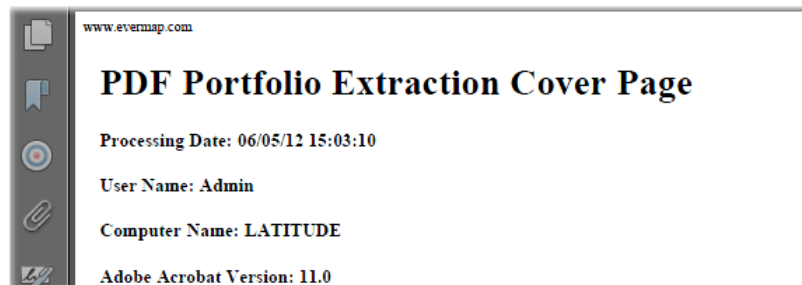


This dialog provides a way to select the options that affect how certain non-PDF files are converted into PDF format. Click the “Select Folder...” button and select a folder on your computer where you want to place copies of the file attachments that did not convert into PDF format or PDF files that cannot be inserted into output due to security restrictions. This option is useful for detailed inspection of the conversion results and provides a way to preserve documents that cannot be converted into PDF format. Press “OK” to close the “File Conversion Options” dialog.

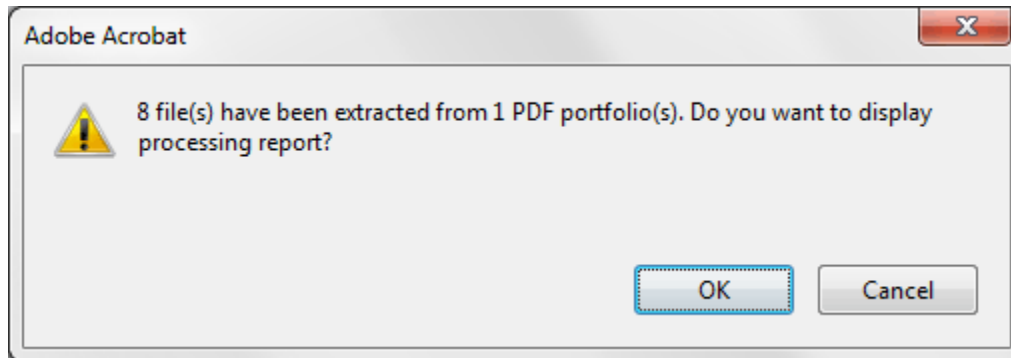
Now, everything is ready to actually run a conversion from a PDF Portfolio into a single, regular PDF document. Press “OK” to start processing:



Please note that the conversion process may take a considerable amount of time depending on the size of the input portfolio. It is generally a good idea to process smaller portfolios. The “PDF Portfolio Extraction Cover Page” document is automatically created for each job and displayed on the screen. The standard Acrobat progress dialog shows the progress at the bottom-right corner of the screen.



Once the processing is completed, a report message is going to appear on the screen asking if you want to display a processing report. Click “OK” to display a detailed processing report in your default web browser (this report is in HTML format):



Here is a sample processing report that lists processed files (there is also spreadsheet ready version of the report in CSV file format). The report lists the file name, description (from a metadata field), creation and modification dates, file size in bytes, number of attachments, and MD5 hash value:

PDF Portfolio Extraction Report

file:///C:/Data/portfolio/PDF%20Portfolio%20Extraction%20Report.htm

AutoPortfolio™ Processing Report

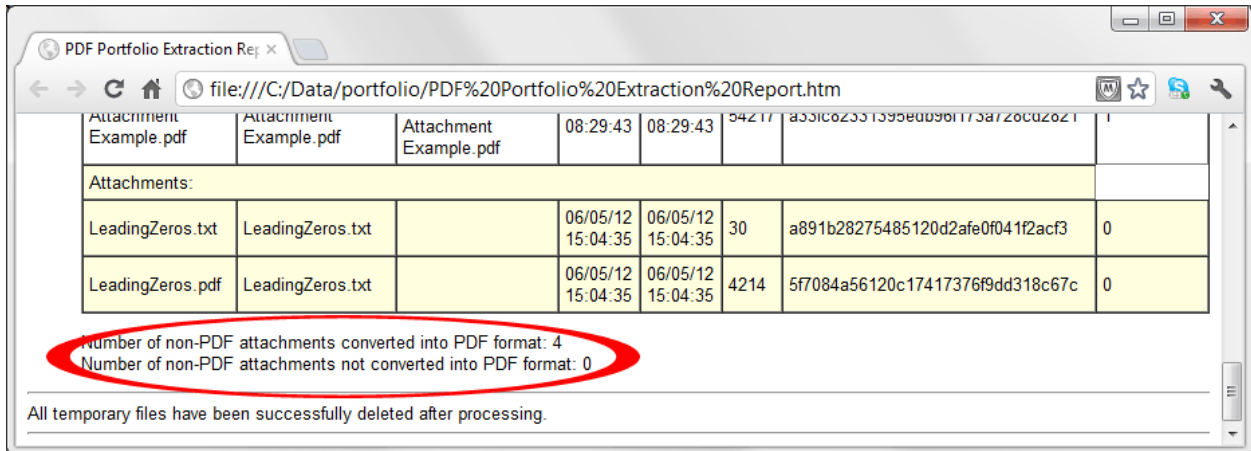
Total Number of PDF Portfolios processed: 1
Output folder: C:\Data\portfolio
Spreadsheet-ready (CSV text file) version of this report: <C:\Data\portfolio\PDF Portfolio Extraction Report.csv>

- Input PDF Package / File: <C:\EverMap PDF\Test Documents\AutoPortfolio\Emails With Different Attachments.pdf>

Total number of extracted files (first level files plus attachments): 8
 Number of extracted files: 4
 Number of extracted attachments: 4

Filename	Original Filename	Description	Created	Modified	File Size	MD5 Hash	Num Attachments
THIS IS A PORTFOLIO TEST.pdf	THIS IS A PORTFOLIO TEST.pdf	"EverMap LLC" 9/12/2011 THIS IS A PORTFOLIO TEST.pdf	09/13/11 08:29:43	09/13/11 08:29:43	59038	edca43ddeb35b3809a5de54790dfe22b	1
Attachments:							
Multilinetext.xlsx	Multilinetext.xlsx		06/05/12 15:03:32	06/05/12 15:03:32	8334	b0b1a51e23bb714ac84967e7363c7cc9	0
Multilinetext.pdf	Multilinetext.xlsx		06/05/12 15:04:18	06/05/12 15:04:18	26351	e76d30360bf6ccf2f99bcb745df6bdb5	0
Test for JPG attachments.pdf	Test for JPG attachments.pdf	"EverMap LLC" 9/13/2011 Test for JPG attachments.pdf	09/13/11 08:29:43	09/13/11 08:29:43	56250	5fbfb00ecf2e253a60691365645bb24b	1

It is a good idea to inspect the report and see if there are any file attachments that were not converted into a PDF file format. You would see a red line in the table for every file attachment that failed to convert. Scroll down to the end of the “AutoPortfolio Process Report” to see the total count of non-PDF file attachments that were converted and not-converted into a PDF format.

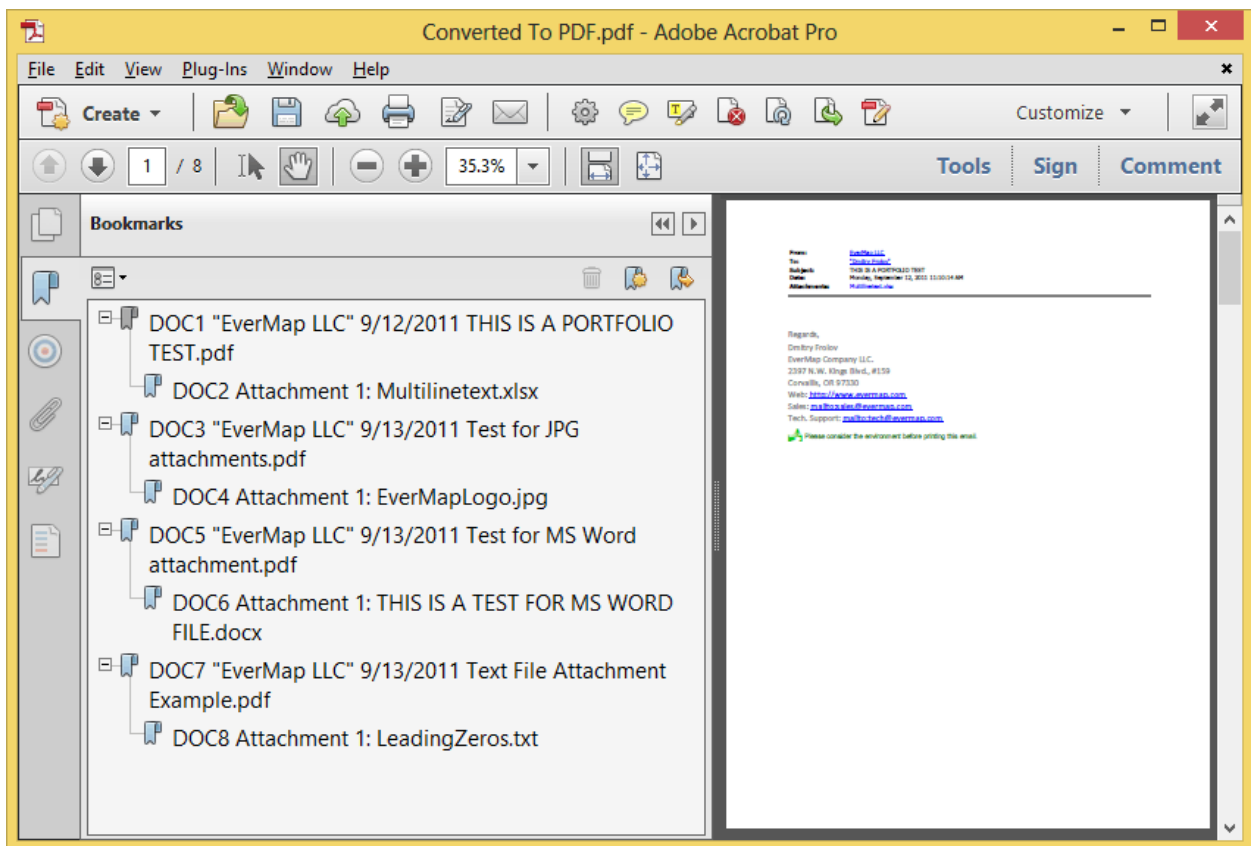


Attachment	Attachment	Attachment	08:29:43	08:29:43	54Z 17	8331c62331395e0090173a720c02021	1
Example.pdf	Example.pdf	Example.pdf					
Attachments:							
LeadingZeros.txt	LeadingZeros.txt		06/05/12 15:04:35	06/05/12 15:04:35	30	a891b28275485120d2afe0f041f2acf3	0
LeadingZeros.pdf	LeadingZeros.txt		06/05/12 15:04:35	06/05/12 15:04:35	4214	5f7084a56120c17417376f9dd318c67c	0

Number of non-PDF attachments converted into PDF format: 4
Number of non-PDF attachments not converted into PDF format: 0

All temporary files have been successfully deleted after processing.

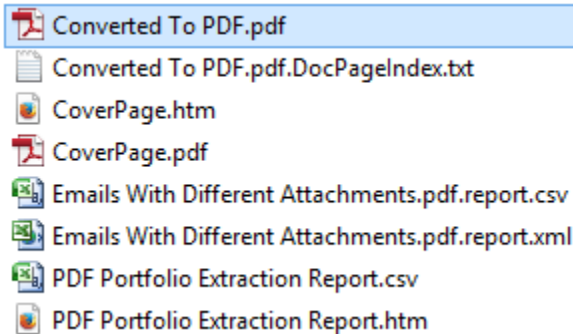
Now, let's look at the output PDF document that is going to be automatically open in Adobe Acrobat once processing is completed:



All individual PDF portfolio entries (PDF files) are now stored in a single PDF document that can be processed using standard Adobe Acrobat tools. Every file (email in the case of email portfolios) is bookmarked with child bookmarks pointing to the corresponding file attachments. The "Description" metadata field and unique document ID is used for the bookmark title. Click on the bookmark to automatically display a corresponding document or attachment. The output document is ready now to

be Bates-stamped and/or printed. The output document is automatically named “Converted To PDF.pdf”. Use the “File > Save As...” menu to save this document under a different filename or location.

Now, let’s look at the output folder and list all the files that are created after the processing. Here is an output from a sample PDF portfolio named *Emails With Different Attachments.pdf*:



The output folder contains the following output files:

- *Converted To PDF.pdf* – output PDF file that contains all entries from the input portfolio as a single, “flat” PDF document.
- *Converted To PDF.pdf.DocPageIndex.txt* - a tab-delimited plain text file that maps document ID to a page number in the output document.
- *CoverPage.pdf* – cover page document that contains information about who and when ran the processing.
- *CoverPage.htm* – same as the above, but in HTML format
- *NamedOfTheInputPortfolio.pdf.report.csv* – Plain text CSV spreadsheet (can be opened directly in Microsoft Excel) that contains all metadata for all PDF Portfolio entries. In the case of the email portfolio, this file will contain “To”, “From”, “Subject” and other similar fields. This file also lists document IDs that used for bookmarking and in the page index file.
- *PDF Portfolio Extraction Report.csv* – plain text CSV spreadsheet (can be opened directly in Microsoft Excel) that contains a processing report listing all files that have been processed and converted.
- *PDF Portfolio Extraction Report.htm* – HTML version of the processing report.

Here is an example of *Converted To PDF.pdf.DocPageIndex.txt* page index file:

```
DOC1  "EverMap LLC" 9/12/2011 THIS IS A PORTFOLIO TEST.pdf 1
DOC2  Attachment 1: Multilinetext.xlsx 2
DOC3  "EverMap LLC" 9/13/2011 Test for JPG attachments.pdf 3
DOC4  Attachment 1: EverMapLogo.jpg 4
DOC5  "EverMap LLC" 9/13/2011 Test for MS Word attachment.pdf 5
DOC6  Attachment 1: THIS IS A TEST FOR MS WORD FILE.docx 6
DOC7  "EverMap LLC" 9/13/2011 Text File Attachment Example.pdf 7
DOC8  Attachment 1: LeadingZeros.txt 8
```



IMPORTANT: Please note that all file format conversions are done by Adobe Acrobat itself using currently selected preferences. Select “File > Preferences...” from the menu to edit the preferences. Select the “Convert to PDF” category, then select a file format to edit and click the “Edit Settings...” button. You may want to make changes to the default settings for Microsoft Excel conversion to allow processing of all worksheets in the file. By default, Adobe Acrobat will only convert the first worksheet in the spreadsheet.

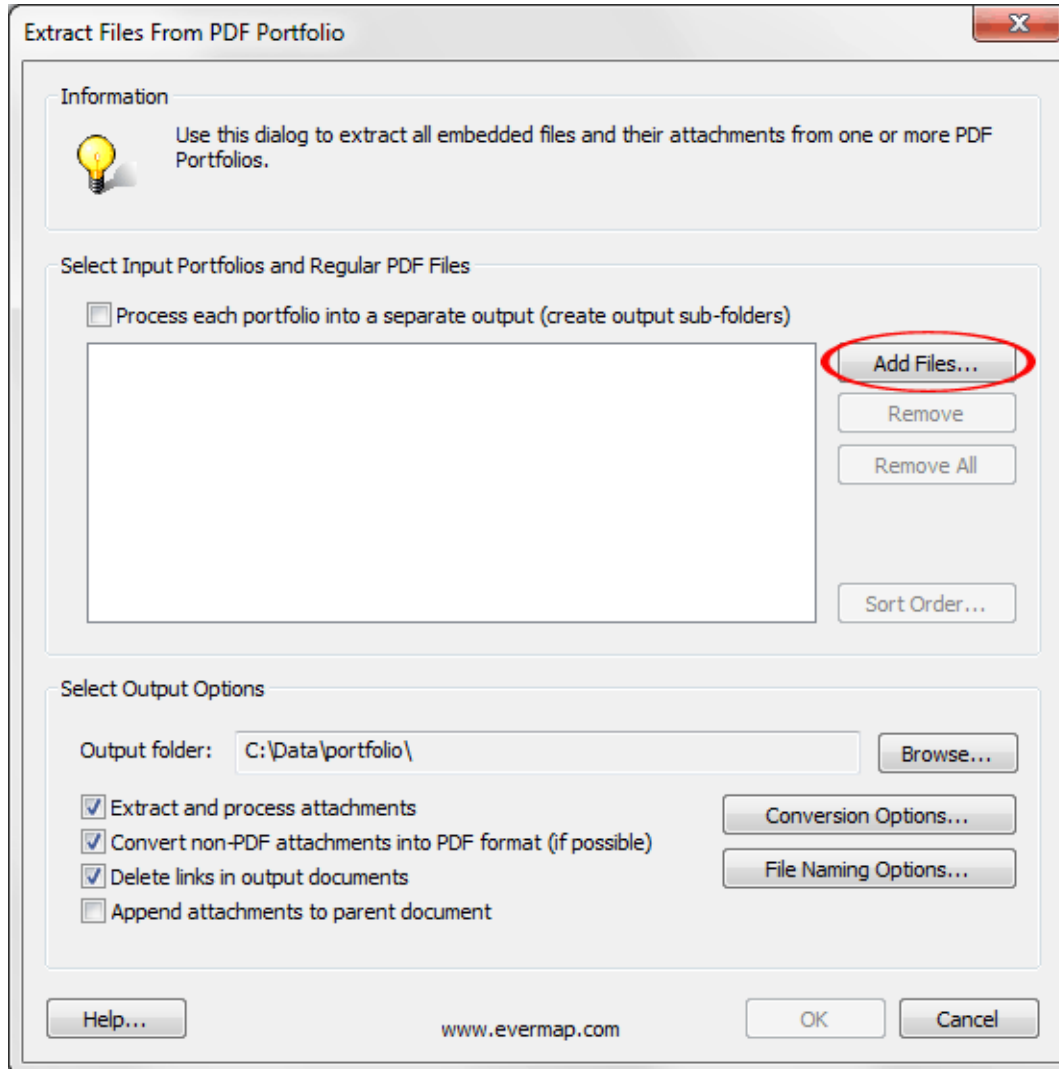


There are certain kinds of PDF and Microsoft Office documents that cannot be merged by Adobe Acrobat into a single PDF document:

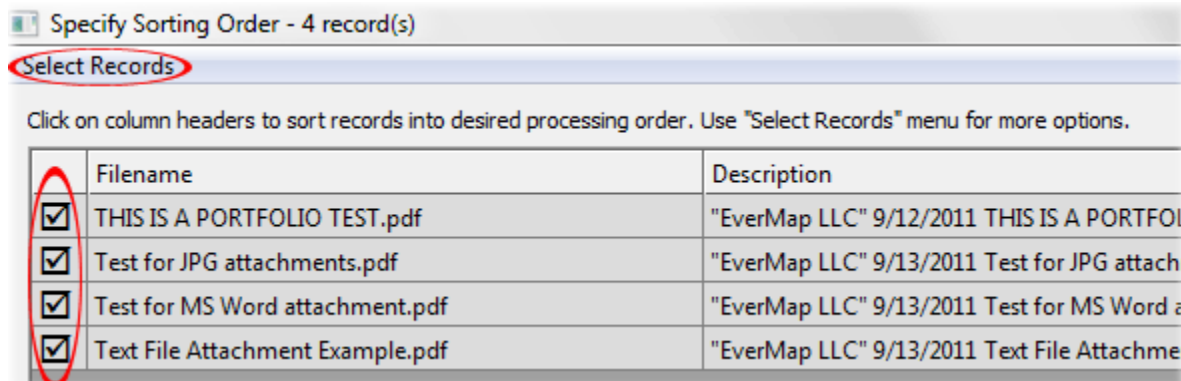
- Password-protected PDF documents (you need to remove passwords before merging)
- Password-protected or encrypted MS Office files
- PDF documents with security restrictions that prevent document merging
- PDF documents with security restrictions that do not allow printing
- MS Office files with access/editing restrictions
- Certain PDF forms created by Adobe LiveCycle Designer form editor

Extract All Files from PDF Portfolio

Select the “Plug-ins > AutoPortfolio Plug-in > Extract Files from Portfolio” menu to extract all files (including attachments) from one or more PDF Portfolios. Do not open a PDF Portfolio directly in Adobe Acrobat; otherwise Acrobat will automatically disable most menus including the “Plug-ins”.

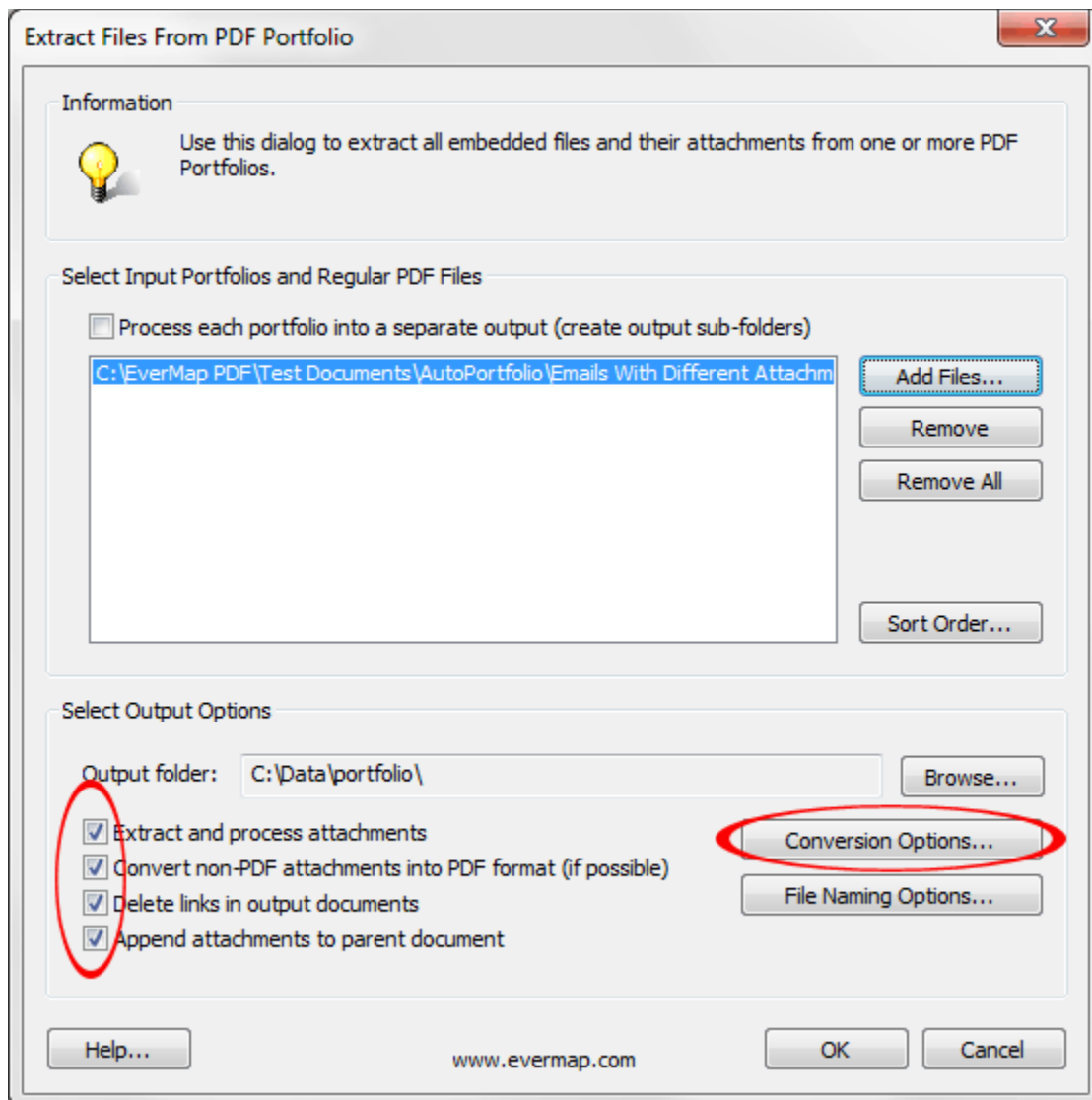


Click the “Add Files...” button to select the input PDF portfolio for processing. The “Specify Sorting Order” dialog will appear on screen once the input file is selected:

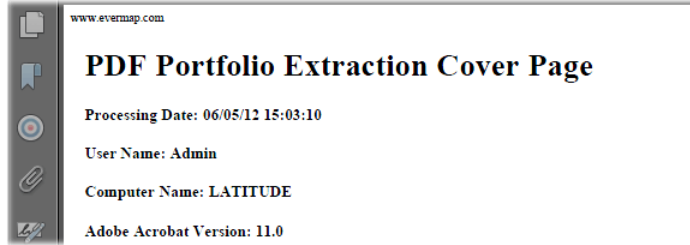


This dialog allows selecting only a part of the PDF Portfolio for processing either by checking a box in front of every record or by performing a text search. If you want to process the entire portfolio, then simply press the “OK” button located in the bottom-right corner of the screen.

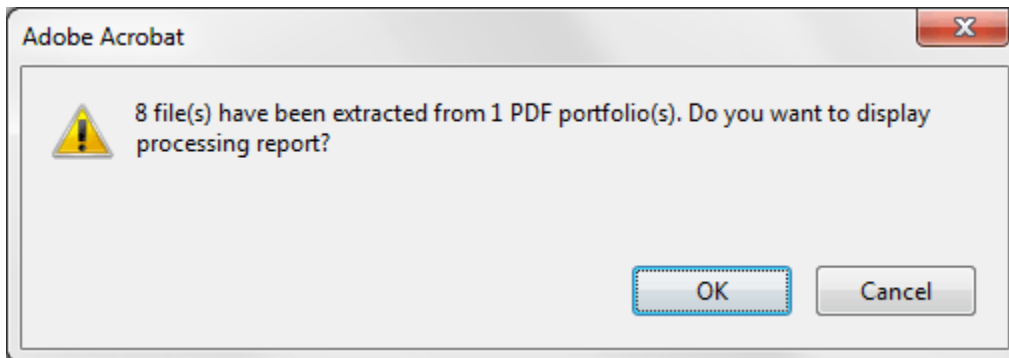
Select an output folder by pressing the “Browse...” button. Check “Extract and process attachments” if you want to extract all attachment files as well. Check the “Convert non-PDF attachments into PDF format” option if you want to convert all supported non-PDF files into PDF format. Check “Append attachments to parent documents” if you want to append all file attachments at the end of their parent PDF documents. This way each portfolio entry (“email” in case of email-based portfolios) contains all its attachments in a single PDF file.



Press “OK” to start extracting files from the input PDF portfolio. The “PDF Portfolio Extraction Cover Page” document is automatically created for each job and displayed on the screen. The standard Acrobat progress dialog shows the progress at the bottom-right corner of the screen.



Once processing is completed, the report message will appear on the screen asking if you want to display a processing report. Click "OK" to display a detailed processing report in your default web browser (this report is in HTML format):



Here is a sample processing report that lists processed files (there is also a spreadsheet ready version of the report in CSV file format). The report lists file name, description (from a metadata field), creation and modification dates, file size in bytes, number of attachments, and MD5 hash value.

PDF Portfolio Extraction Report

file:///C:/Data/portfolio/PDF%20Portfolio%20Extraction%20Report.htm

AutoPortfolio™ Processing Report

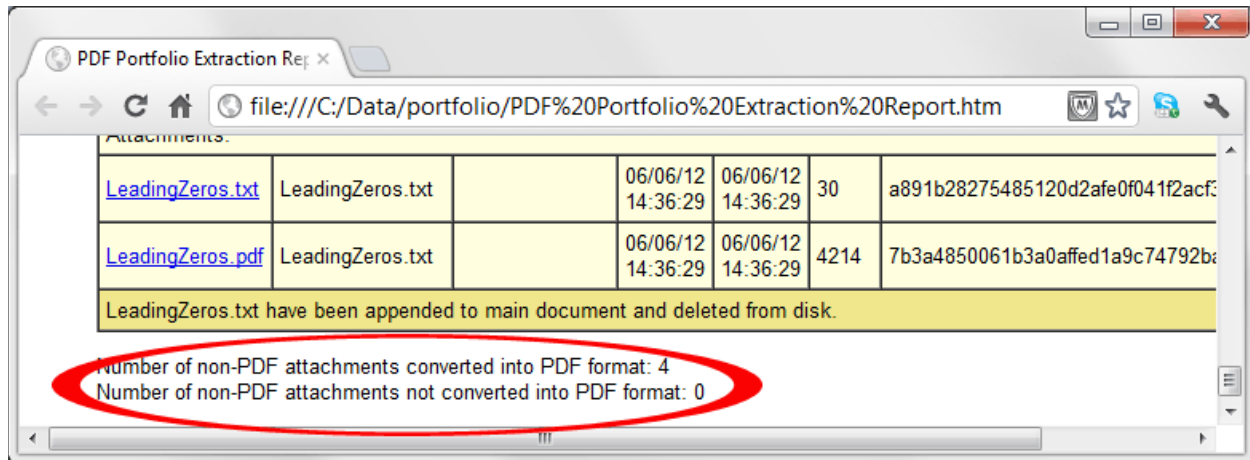
Total Number of PDF Portfolios processed: 1
Output folder: C:\Data\portfolio
Spreadsheet-ready (CSV text file) version of this report: [C:\Data\portfolio\PDF Portfolio Extraction Report.csv](#)
Load file (list of all processed files in text format): [C:\Data\portfolio\LoadFile.txt](#)

- Input PDF Package / File: [C:\EverMap PDF\Test Documents\AutoPortfolio\Emails With Different Attachments.pdf](#)

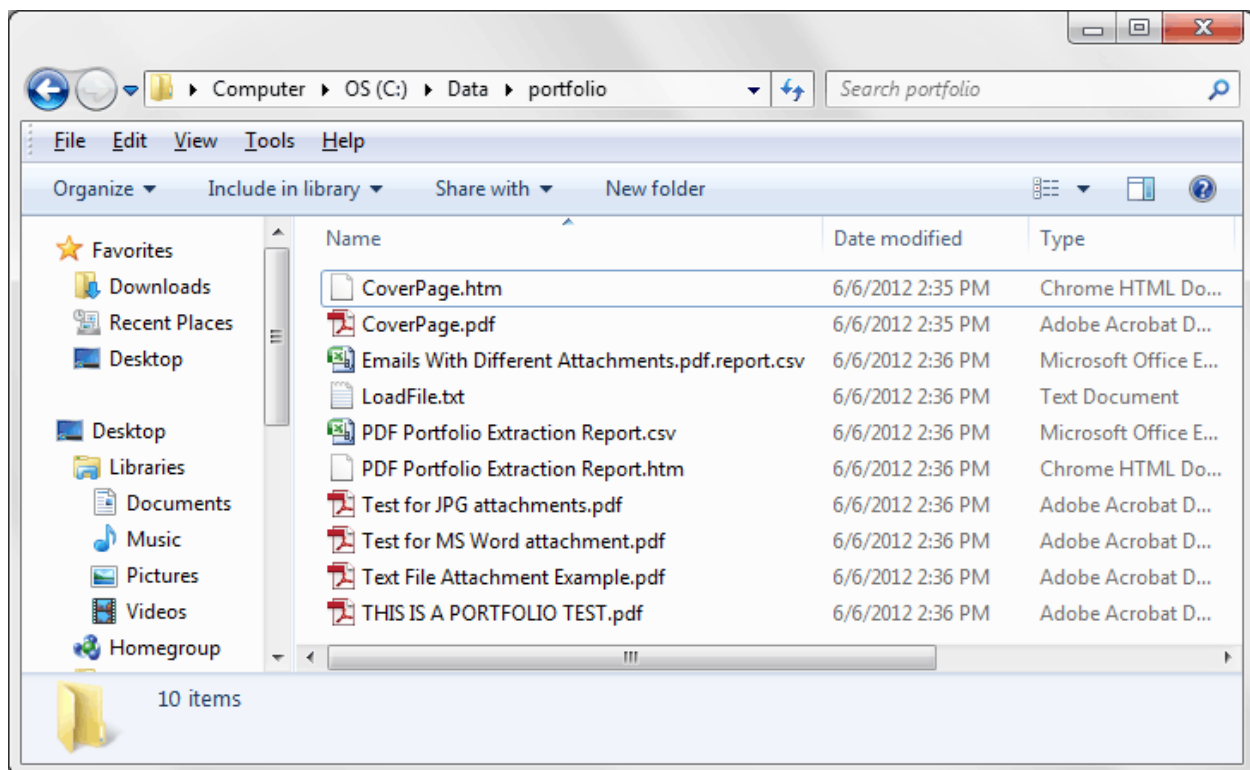
Total number of extracted files (first level files plus attachments): 8
 Number of extracted files: 4
 Number of extracted attachments: 4

Filename	Original Filename	Description	Created	Modified	File Size	MD5 Hash	Num Attachments
THIS IS A PORTFOLIO TEST.pdf	THIS IS A PORTFOLIO TEST.pdf	"EverMap LLC" 9/12/2011 THIS IS A PORTFOLIO TEST.pdf	09/13/11 08:29:43	09/13/11 08:29:43	59038	edca43ddeb35b3809a5de54790dfe22b	1
Attachments:							
Multilinetext.xlsx	Multilinetext.xlsx		06/06/12 14:35:31	06/06/12 14:35:31	8334	b0b1a51e23bb714ac84967e7363c7cc9	0
Multilinetext.pdf	Multilinetext.xlsx		06/06/12 14:36:12	06/06/12 14:36:12	26350	6fe766ed1c0c931276493807c941050f	0
Multilinetext.xlsx have been appended to main document and deleted from disk.							

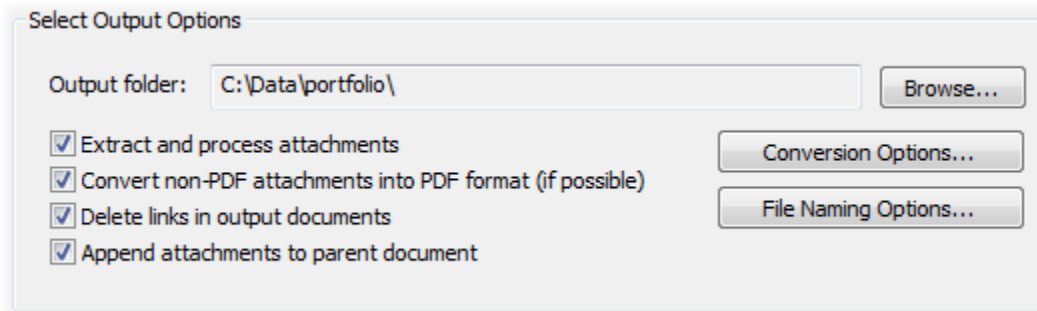
The processing report lists every file and attachment that was processed and contains separate records for the attachments that were converted into PDF format and optionally appended to the parent document. It is a good idea to inspect the report and see if there are any file attachments that were not converted into the PDF file format. You would see a red line in the table for every file attachment that failed to convert. Scroll down to the end of the “AutoPortfolio Process Report” to see the total count of non-PDF file attachments that were converted and not-converted into a PDF format.



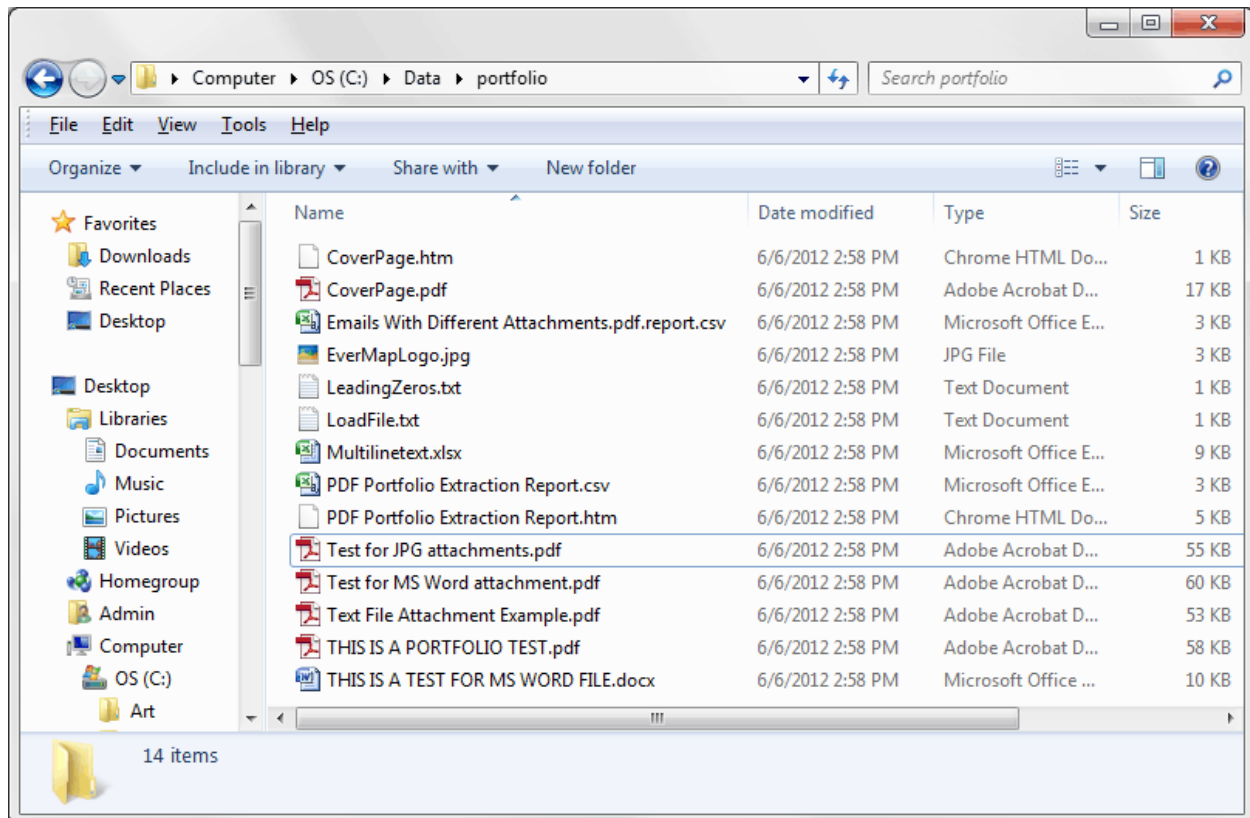
Now, let's take a look at the content of the output folder:



The above snapshot of the output folder is the result of processing when the following processing options were selected:

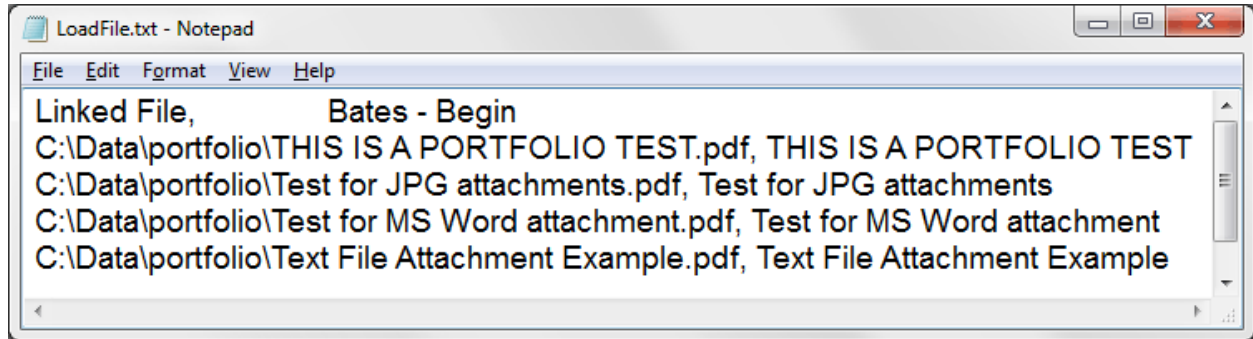


The output folder does not contain attachments in their native file formats since the option “Convert non-PDF attachments into PDF format” was used. The below is another snapshot of the output folder when this option was turned off (as well as the “Append attachments to parent documents”):



The output folder now contains file attachments as well as PDF portfolio items in their native formats. All main level PDF Portfolio items are exported as PDF files (they are already stored in the portfolio in PDF format). The output folder also contains a number of various auxiliary files such as extraction reports (both in CSV and HTML formats), processing cover page, and *.pdf.report.csv file (can be opened directly by Microsoft Excel) that contains metadata for all extracted files. There is also a file called

LoadFile.txt (in CaseMap format) that lists all the extracted top-level files (in the case of email-based portfolios only top level emails):

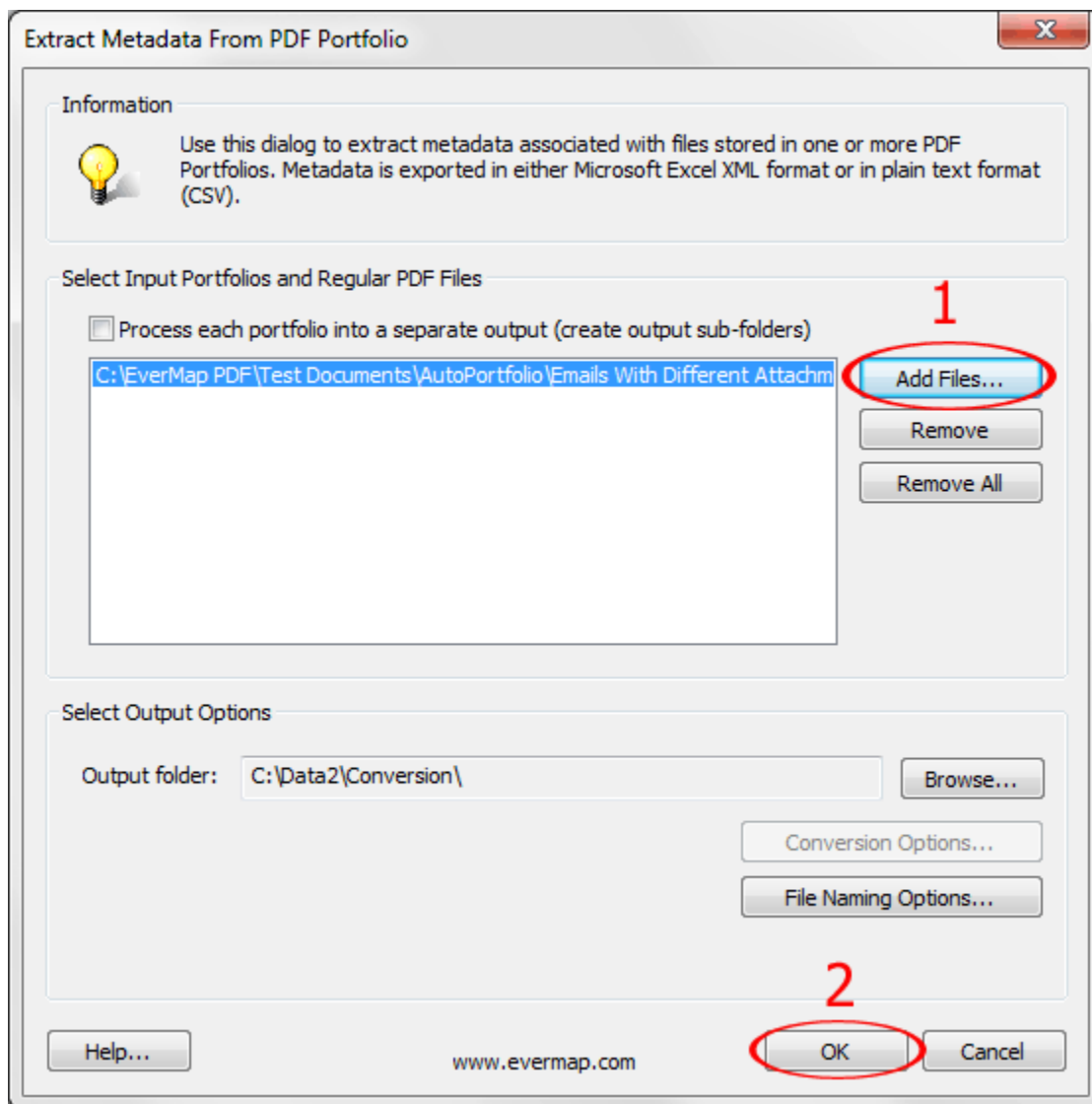


```
LoadFile.txt - Notepad
File Edit Format View Help
Linked File,      Bates - Begin
C:\Data\portfolio\THIS IS A PORTFOLIO TEST.pdf, THIS IS A PORTFOLIO TEST
C:\Data\portfolio\Test for JPG attachments.pdf, Test for JPG attachments
C:\Data\portfolio\Test for MS Word attachment.pdf, Test for MS Word attachment
C:\Data\portfolio\Text File Attachment Example.pdf, Text File Attachment Example
```

If the input PDF Portfolio contains an MSG file (Microsoft Outlook message format) or ZIP archive, then a folder is created for each file that contains files from the corresponding MSG and ZIP archives. If there are nested MSG or ZIP files, the second-level subfolders are automatically created and the files extracted.

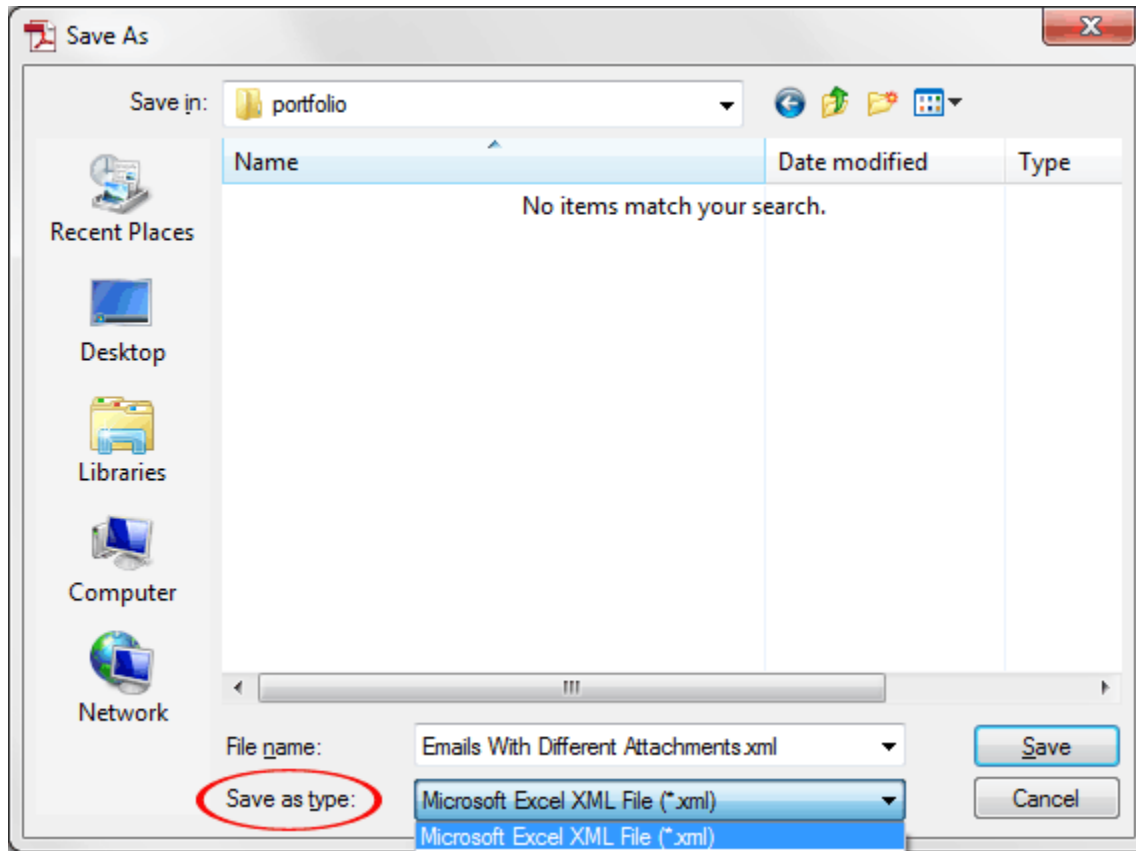
Extracting PDF Portfolio metadata into Microsoft Excel Spreadsheet

The PDF Portfolio is a single storage file that contains a number of files along with the corresponding metadata associated with each file. The actual metadata fields included totally depend on the application that created the particular portfolio. The most common use of PDF Portfolio is to store emails exported from email applications such as Microsoft Outlook or Lotus Notes. In this case, most metadata fields come from the corresponding email's metadata and typically include "To", "From", "Subject" and etc. The AutoPortfolio plug-in can export all this information into an Excel spreadsheet file. The plug-in exports **all fields** that are actually present in the PDF Portfolio along with some additional metadata fields that are computed by the plug-in itself (MD5 hash).



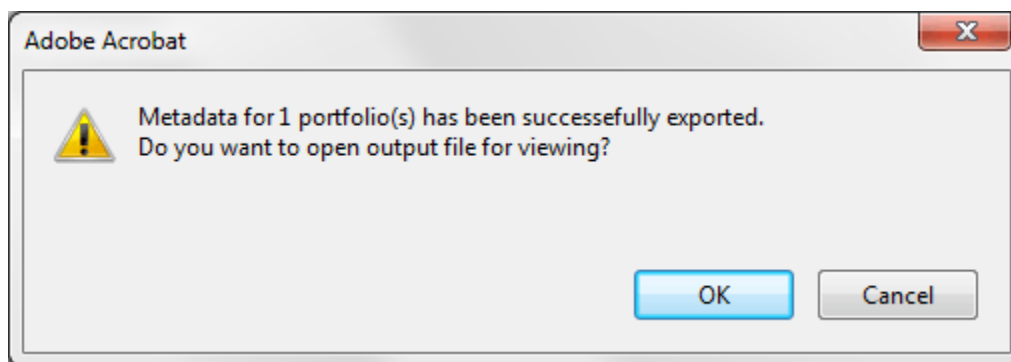
Click the "Add Files..." button to select the input PDF Portfolio and "OK" to start processing.

Use the “Select Type” option menu in the “Save As” dialog to choose a desired output file format:



The plug-in supports two output formats: Microsoft Excel XML format and plain text ASCII *.CSV file format. Both output formats contain the same data.

Press “OK” on the final confirmation dialog to open the output file in Microsoft Excel:

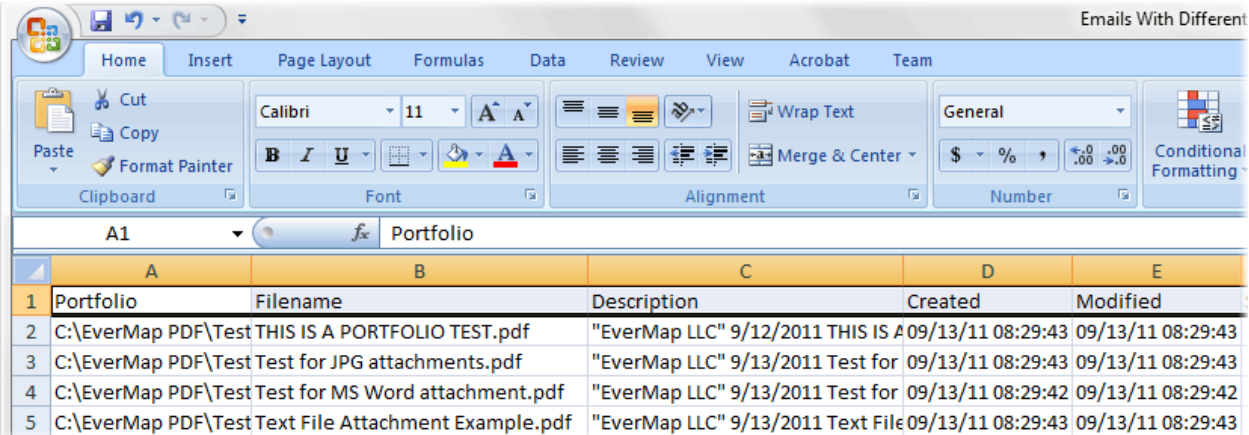


A typical email-based portfolio exported from Microsoft Outlook has the following metadata fields for each email entry:

- “Portfolio” – the full path to the portfolio .
- “Filename” – original filename associated with the email entry.
- “Description” – a composite field that combines “From” field with data and email subject.

- “Created/Modified” – the date this entry was created and modified (this is not the “Sent” date).
- “Size” – the actual size in bytes of the email entry (typically the size of a corresponding PDF file).
- “MD5 Hash” – MD5 hash value for the corresponding file. This number is unique and can be used to compare files.
- “Folder location” – the name of the email folder where this email came from (Personal Folders/Inbox for example).
- “From” – the “From” field of the corresponding email message.
- “To” – the “To” field of the corresponding email message.
- “GUID” – globally unique identifier for the corresponding email message.
- “Date” – the actual date the email was received.
- “Attachments” – the number of attachments in the corresponding email message
- “Subject” – subject field of the corresponding email message.

The above fields are listed as an example only. The actual output depends on the actual metadata fields stored in the specific PDF Portfolio. The plug-in does not look for any pre-defined fields; it exports all fields that are present in the actual file.

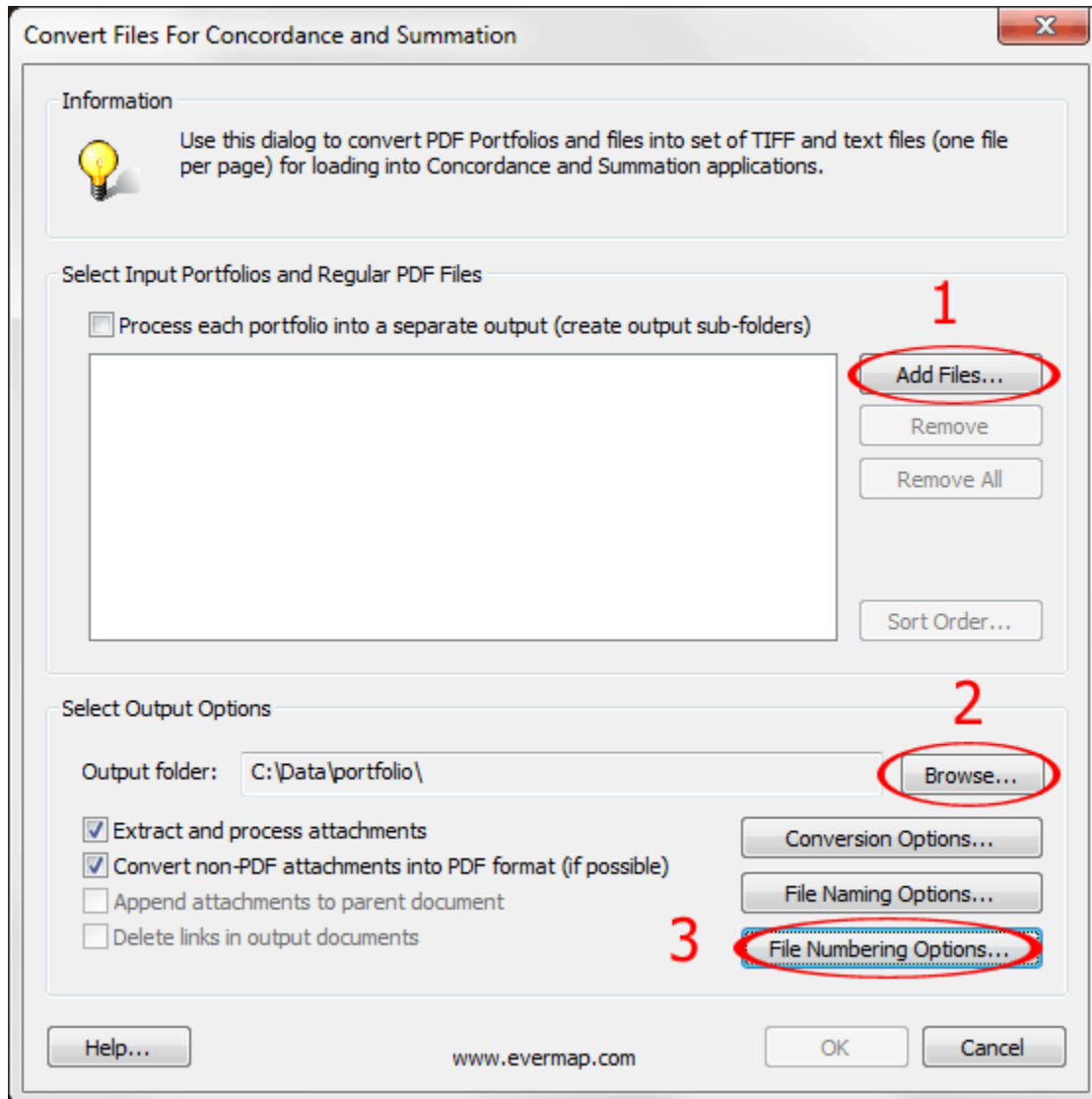


Portfolio	Filename	Description	Created	Modified
C:\EverMap PDF\Test	THIS IS A PORTFOLIO TEST.pdf	"EverMap LLC" 9/12/2011 THIS IS A	09/13/11 08:29:43	09/13/11 08:29:43
C:\EverMap PDF\Test	Test for JPG attachments.pdf	"EverMap LLC" 9/13/2011 Test for	09/13/11 08:29:43	09/13/11 08:29:43
C:\EverMap PDF\Test	Test for MS Word attachment.pdf	"EverMap LLC" 9/13/2011 Test for	09/13/11 08:29:42	09/13/11 08:29:42
C:\EverMap PDF\Test	Text File Attachment Example.pdf	"EverMap LLC" 9/13/2011 Text File	09/13/11 08:29:43	09/13/11 08:29:43

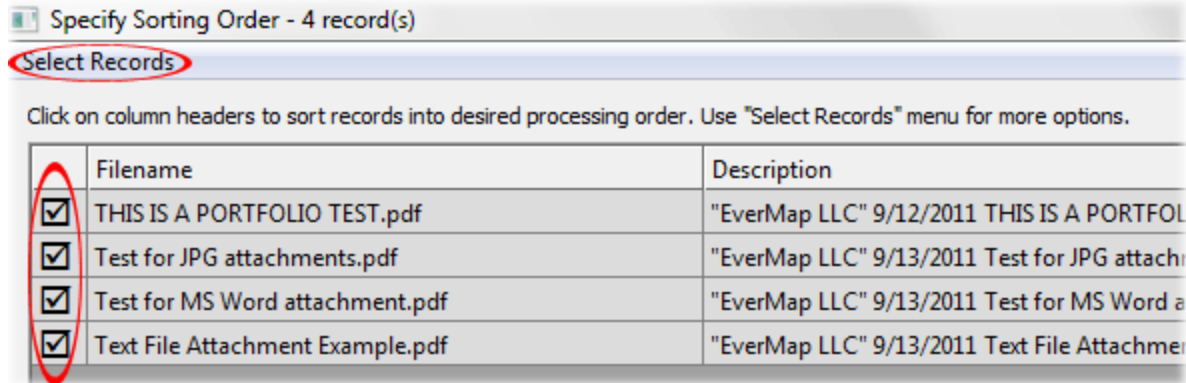
Exporting PDF Portfolio into TIFF and Text format

The plug-in provides the ability to export a PDF Portfolio into a TIFF and Text format that is suitable for importing data into several litigation support systems such as Concordance and Summation. The output is a collection of files – one TIFF image file and one text file for each page of each document in the portfolio.

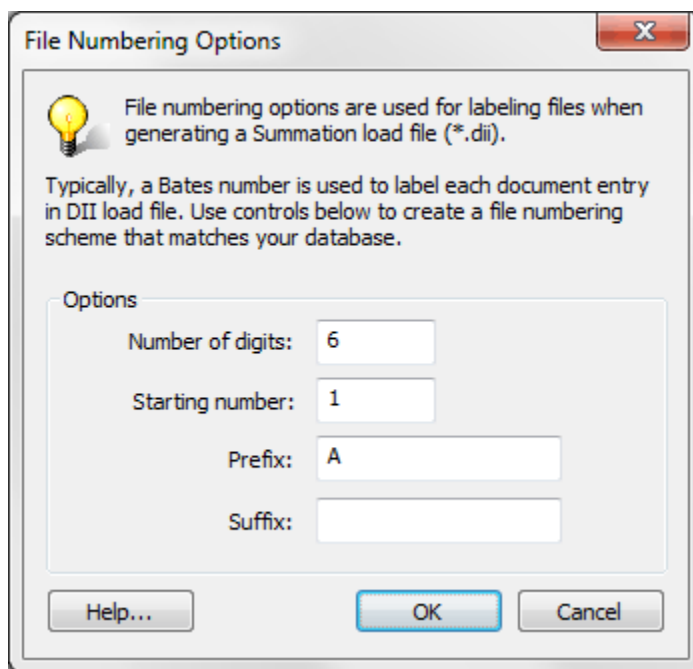
Select “Plug-ins > AutoPortfolio Plug-in > Convert PDF Files for Concordance and Summation (TIFF and Text)” from the menu. Press the “Add Files...” button to select an input PDF Portfolio.



The “Specify Sorting Order” dialog will appear on the screen once the input file is selected. This dialog allows for the selection of only a part of the PDF Portfolio for processing either by checking a box in front of every record or by performing a text search. If you want to select only a specific subset of the portfolio entries for the processing, then use the “Select Records” menu to manipulate the current selection set



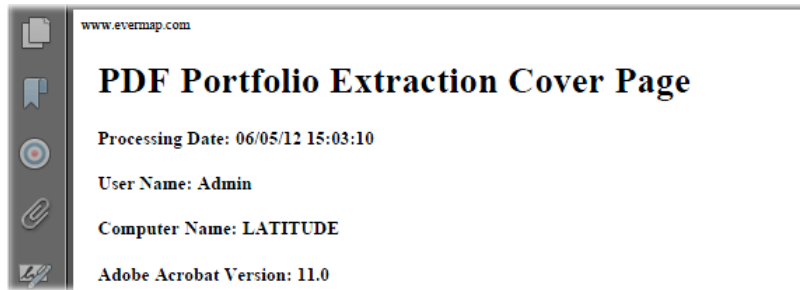
If you want to process the entire portfolio, then simply press "OK" located in the lower-right corner of the screen. Select an output folder by pressing the "Browse..." button. Next, press the "File Numbering Options..." button to choose the desired file numbering settings.



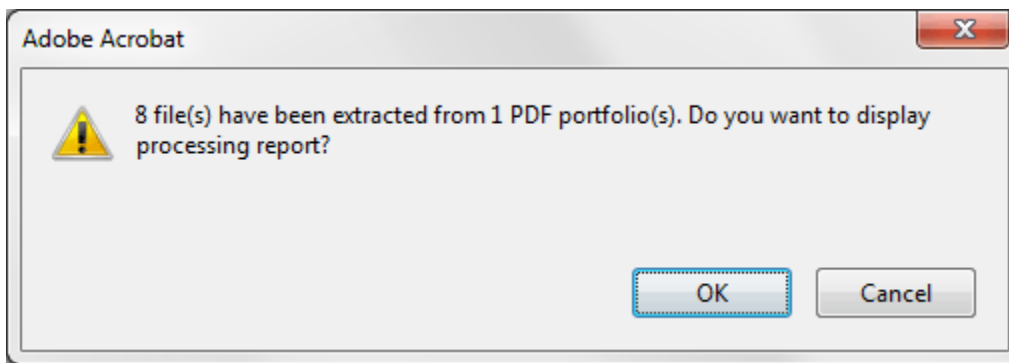
This dialog allows you to configure the plug-in to number output files to match file naming in the Summation database. Please refer to Summation documentation for details on file naming and DII import or press "Help..." to read more about these options.

Now, choose any other output option, then press "OK" to start processing. Please note that the conversion process may take a considerable amount of time depending on the size of the input portfolio. It is generally a good idea to process smaller portfolios. The processing consists of extracting every PDF document out of the portfolio and extracting any existing file attachments. Converting non-PDF attachments into PDF (optional) and then converting all resulting PDF files into TIFF/TEXT format. For every page in a PDF file there are two output files: one TIFF image and one plain text file. The "PDF Portfolio Extraction Cover Page" document is automatically created for each job and displayed on the screen. The standard Acrobat progress dialog shows the progress at the bottom-right corner of

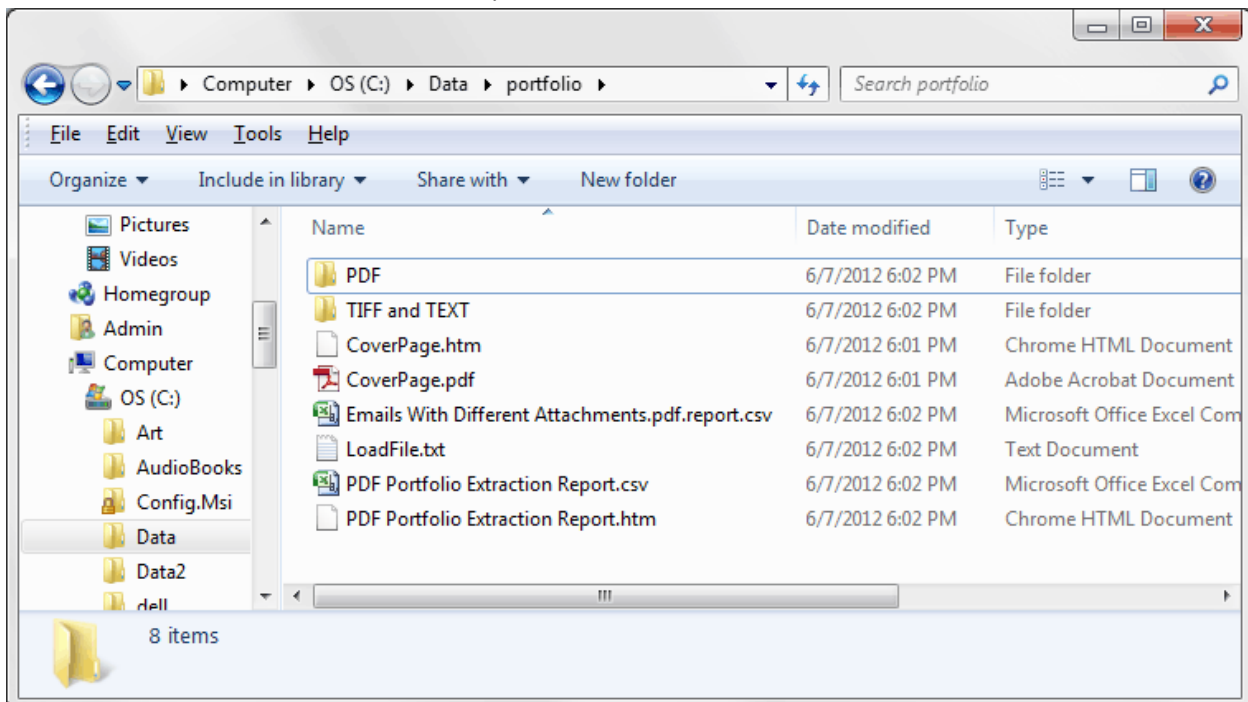
the screen.



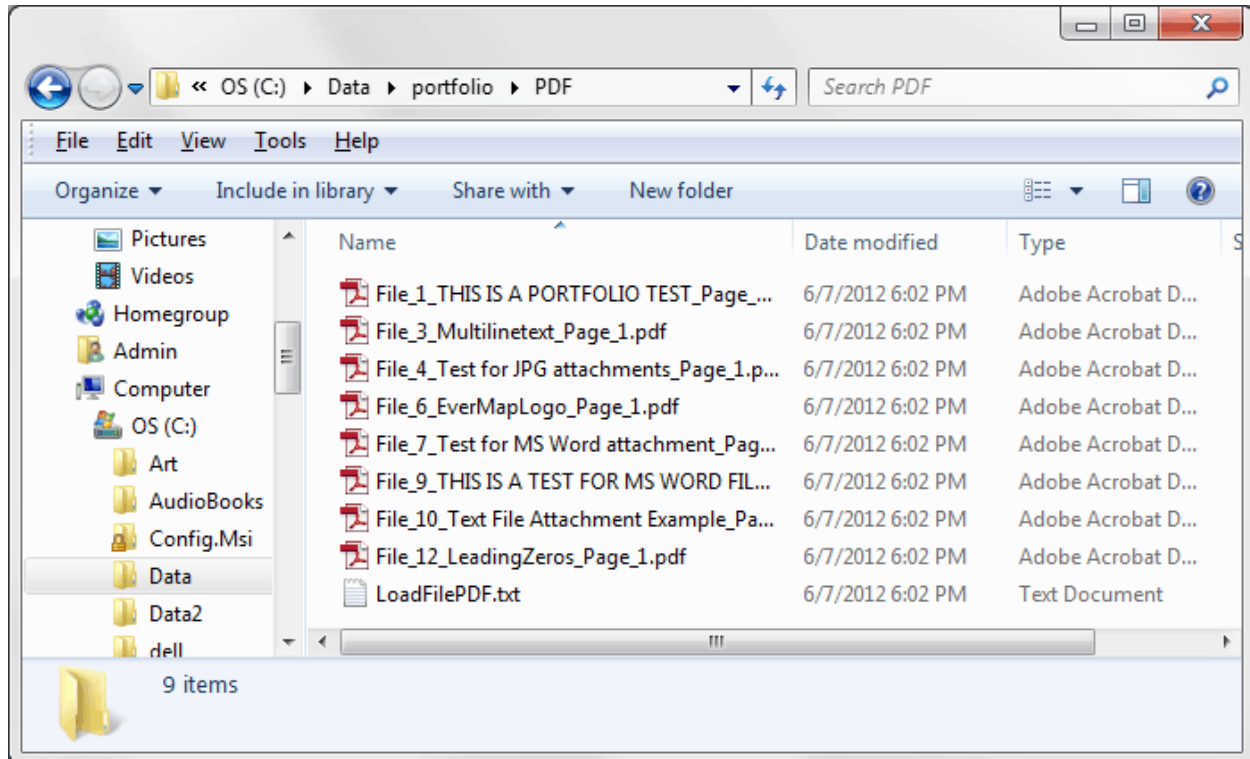
Once processing is completed, the report message will appear on the screen asking if you want to display a processing report. Click "OK" to display a detailed processing report in your default web browser (this report is in HTML format).



Now, let's look at the contents of the output folder:



There are two sub-folders in the output folder: “PDF” and “TIFF and TEXT”. Here is an example of the content for the “PDF” folder:



This folder contains a number of PDF files named File_X_*.pdf. The total number of PDF files in this folder is equal to the total number of pages in all the documents of the input PDF Portfolio. Each PDF file contains only one page.

The “TIFF and TEXT” folder contains:

Number of *.TXT files named File_X_*.txt. There is exactly one text file for each page in the PDF Portfolio.

Number of image *.TIF files named File_X_*.tif.

Here is an example of the content for the “TIFF and TEXT” Folder:

