Part 5 Create Full Text Content using ABBYY FIneReader OCR software

OCR (Optical Character Recognition) software translates the “pictures” of text from your scans into machine-readable text. The quality of this translation depends on the printed text, the quality of scan, and the OCR software processing those scans.

Assessing Text Objects:
There is not much you can do about the quality of the original text printed on pages. If you have any questions about the quality of full text you might be able to generate scans and run a few tests using a trial version of OCR software before investing. **Note:** OCR software will not “read” handwriting with any effectiveness.

Scans
For best results, use a 400dpi greyscale scan for black and white text objects or full colour for colour objects.

OCR Software

ABBYY FineReader Professional OCR software works very well with typed text of most kinds, and can be adjusted to recognize different languages.

NOTE: ABBYY FineReader is now at Version 12. The following screenshots should be used as a guide to using the newer version in terms of workflow, as they do not reflect the updated interface.

Part 5.1 Batch OCR Processing

Before you begin…

You must first have associated multiple files with the record and use those same files for OCR processing (See Part 2).

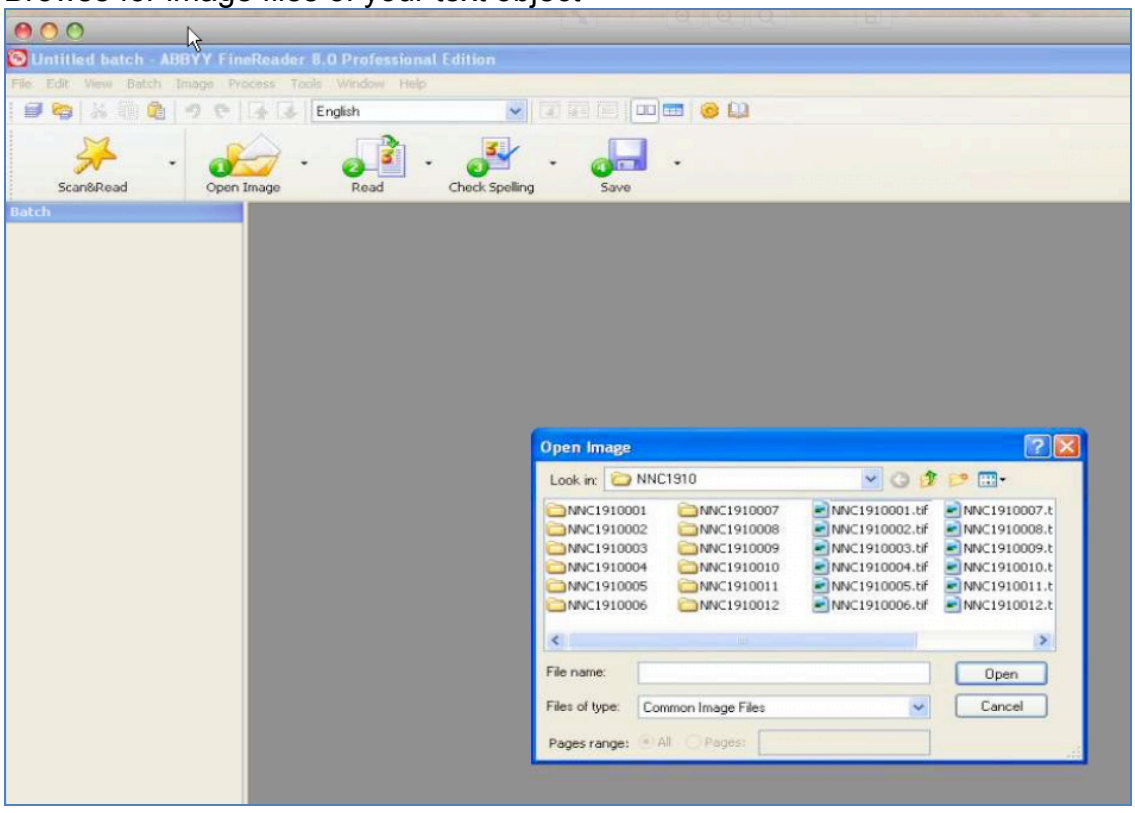
Then…

Open ABBYY FineReader on your desktop

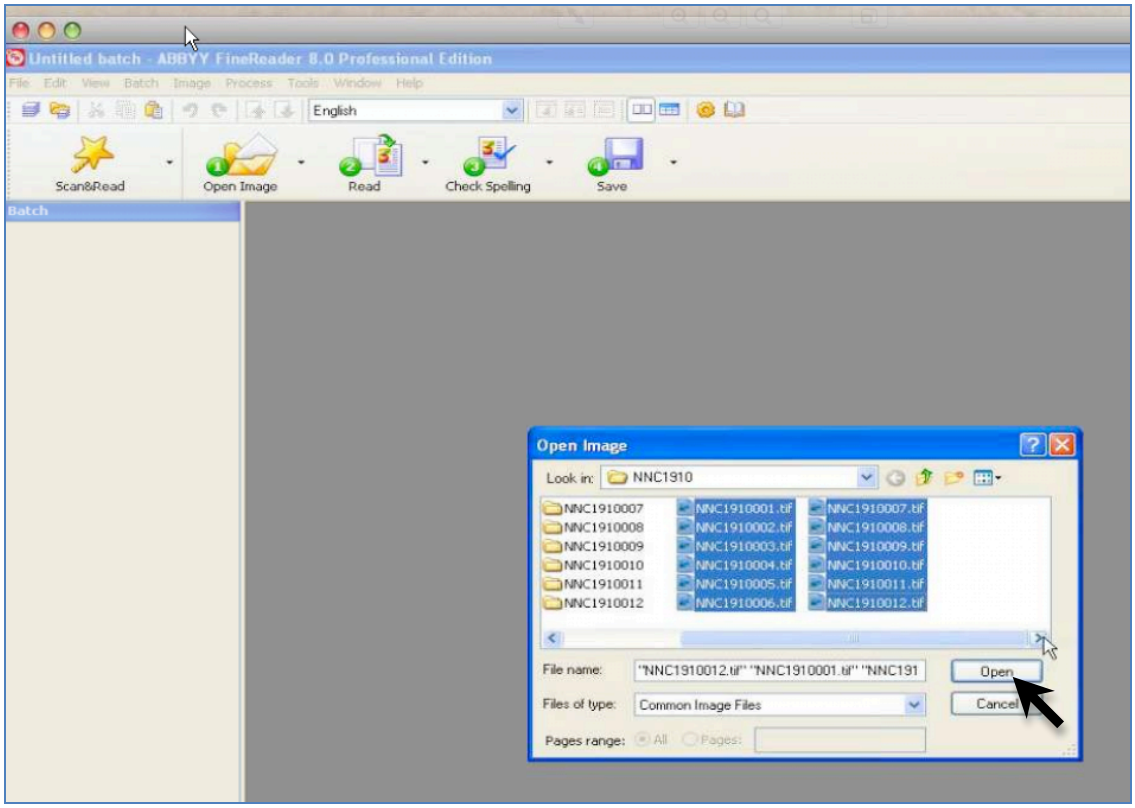
Part 5.1.1 Select Images

Click on the Open & Read button

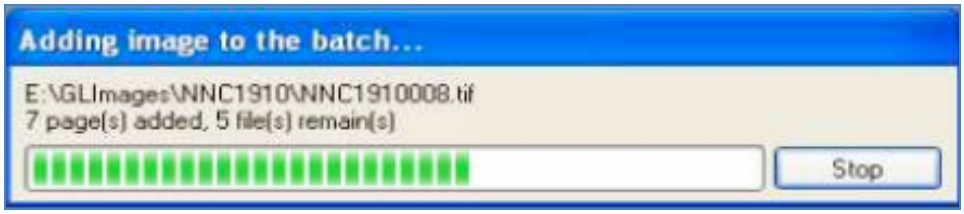
Browse for image files of your text object



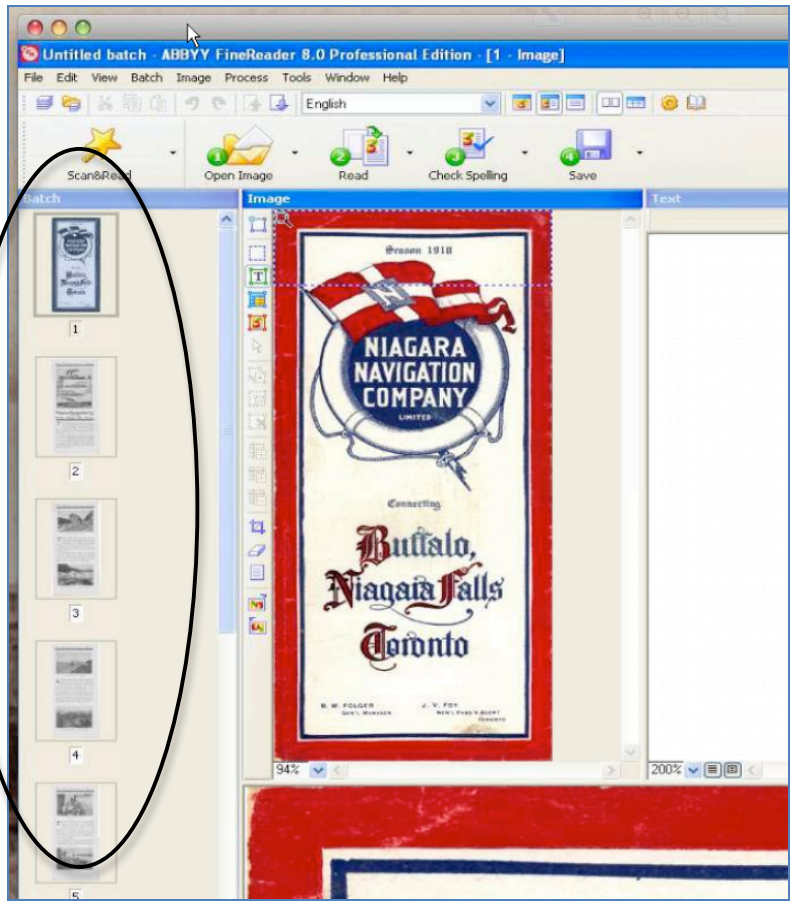
Use your Shift or Command + click keys to select multiple files from your local drive
Click “Open”



Wait while the files are uploaded to ABBYY

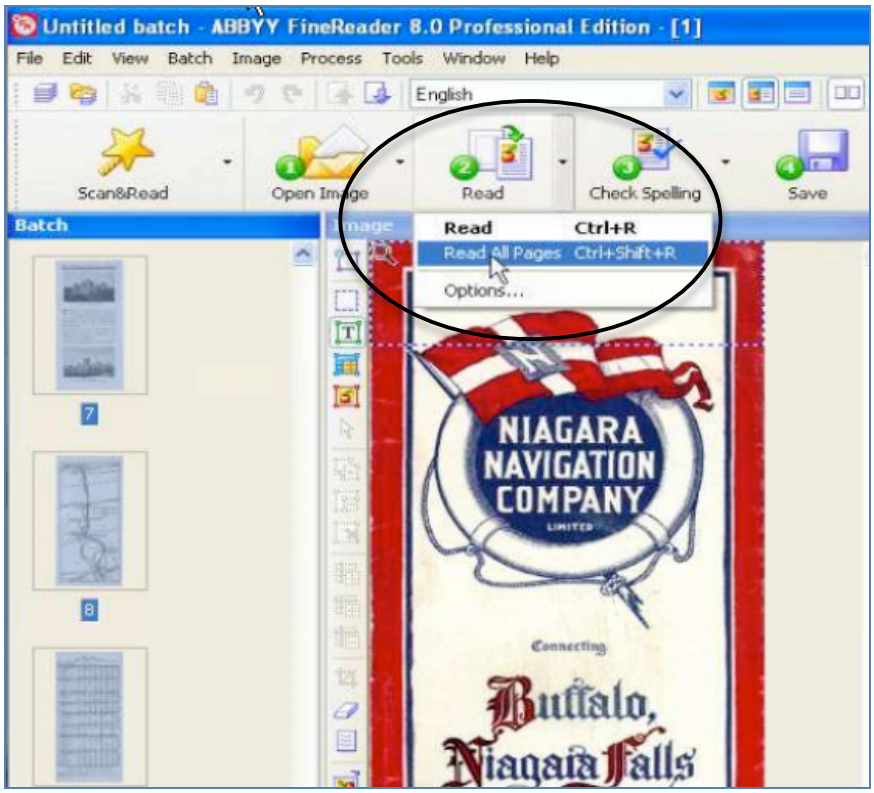


The uploaded files will display as thumbnails

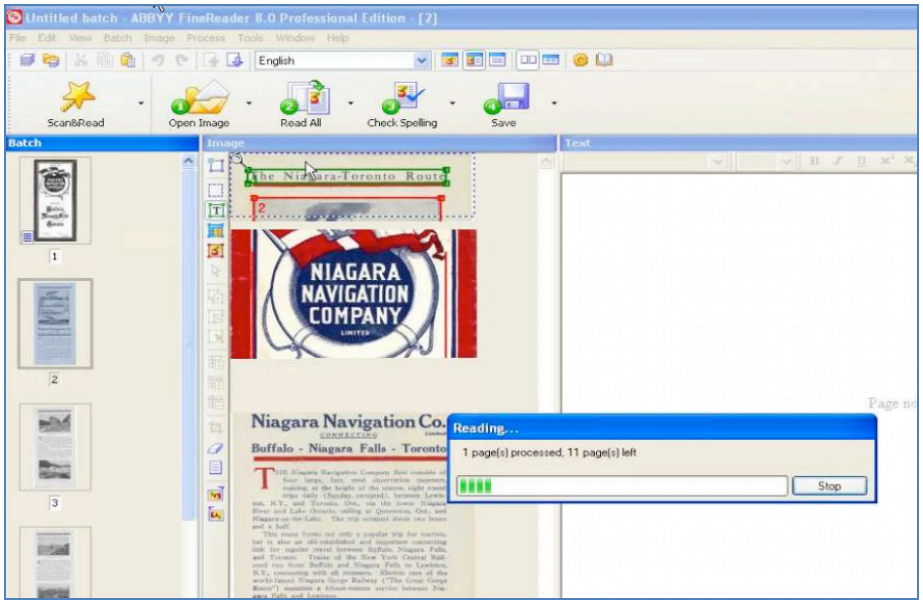


Part 5.1.2 Read Images

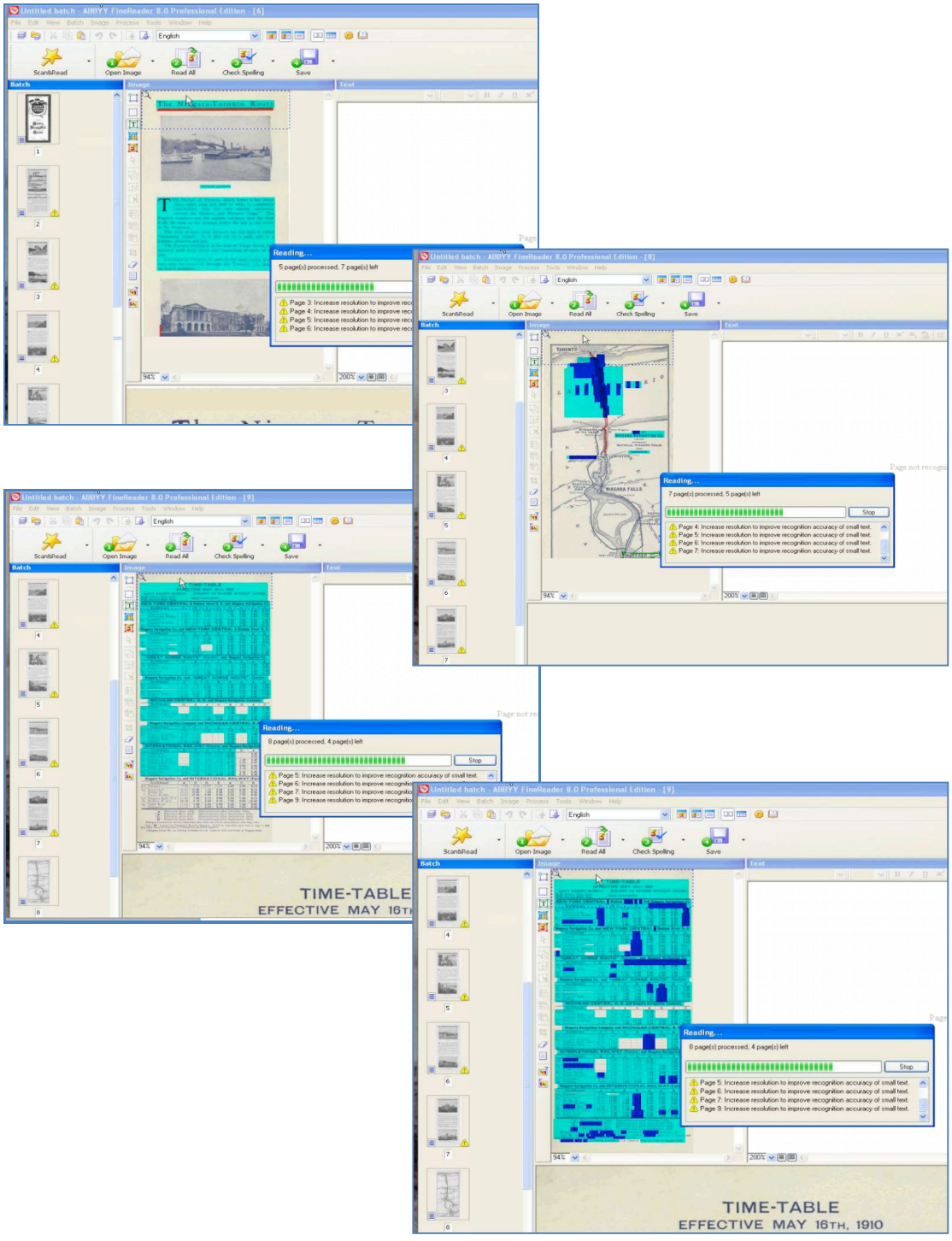
Click on the Read button and select Read All Pages



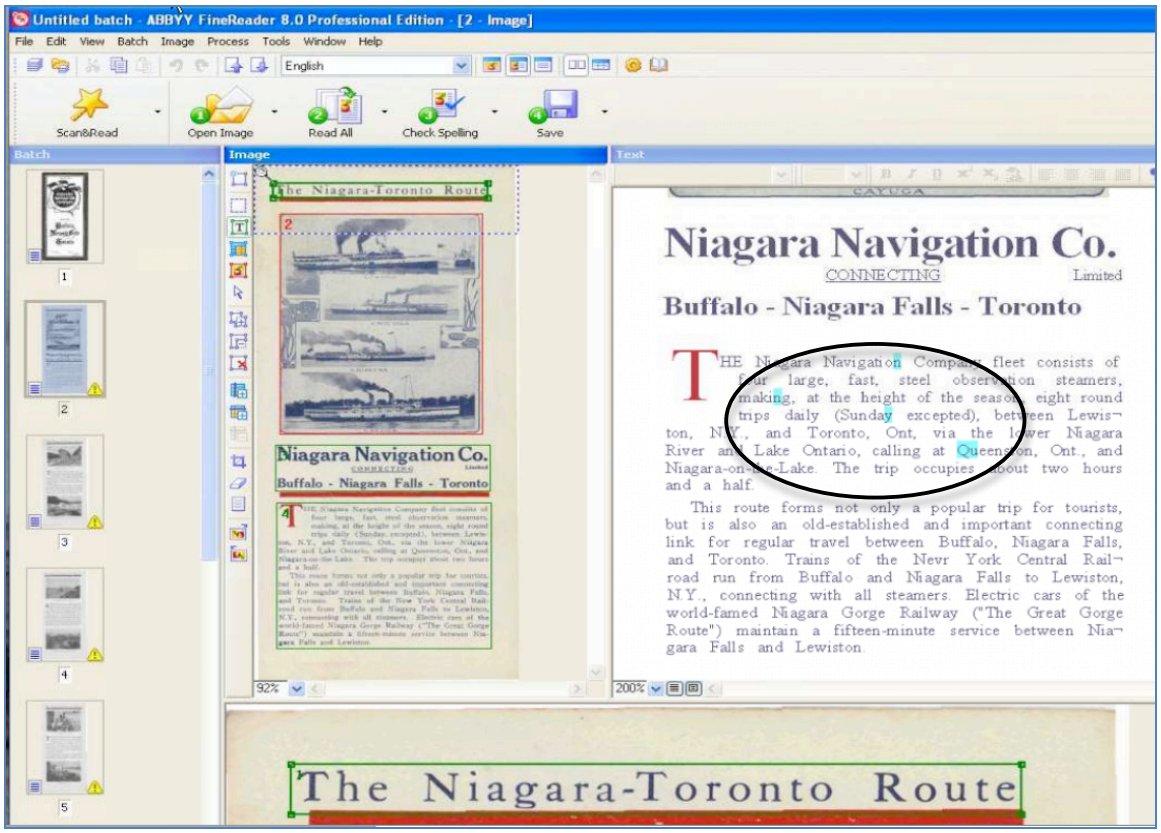
Wait while the application processes the multiple files…



Progress is shown with the highlighting (light blue indicates progress, dark blue indicates uncertainty about character recognition)…

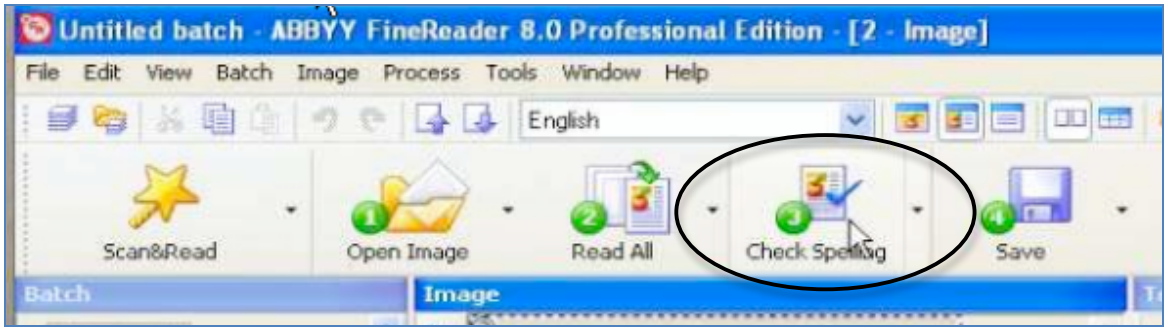


The text from each page will appear on the right hand side. Light blue highlights indicate letters that may be incorrect.

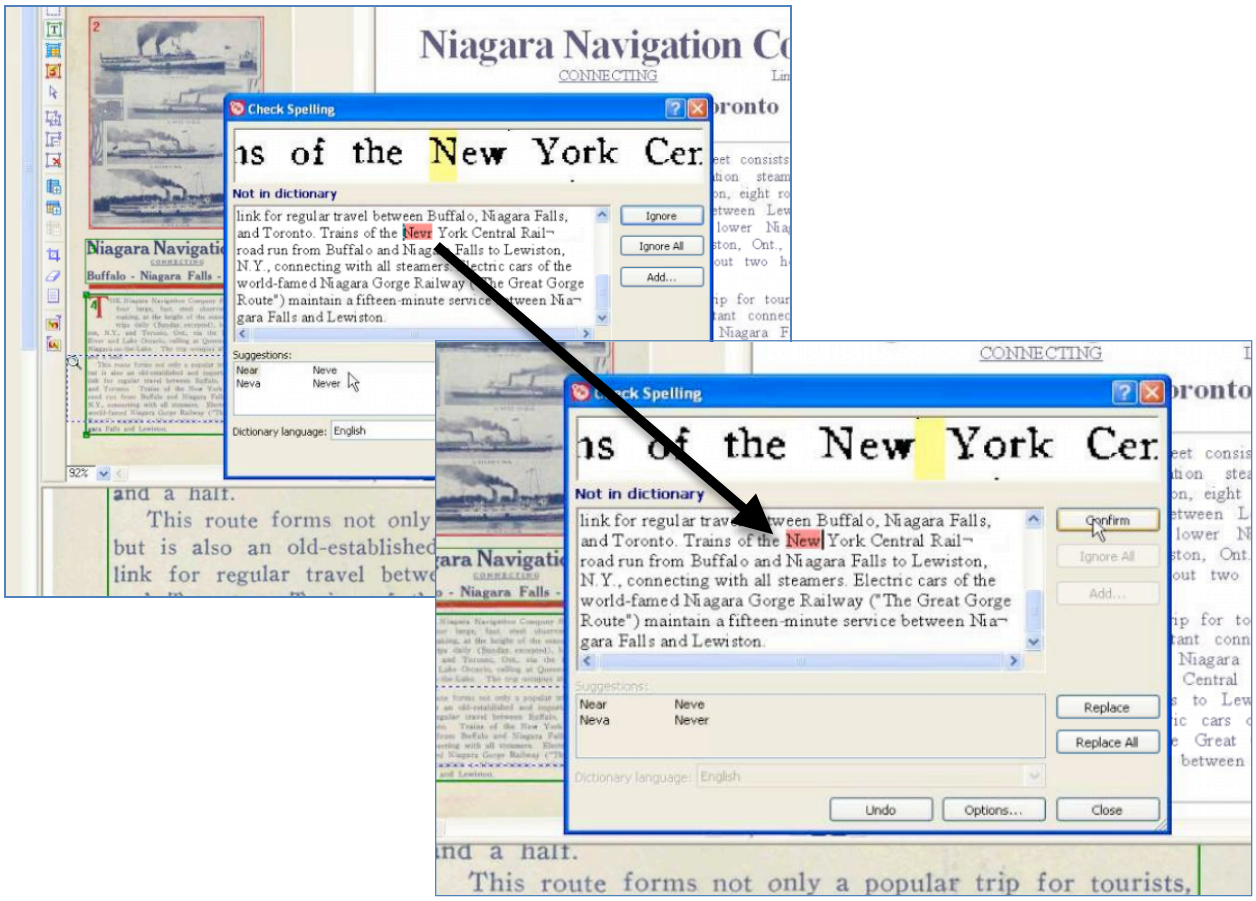


Part 5.1.3 Check Spelling

You can either correct the spelling manually in the right hand panel or use the Check Spelling feature:



The Spell Check window will pop up and highlight incorrect words as well as spelling correction options.



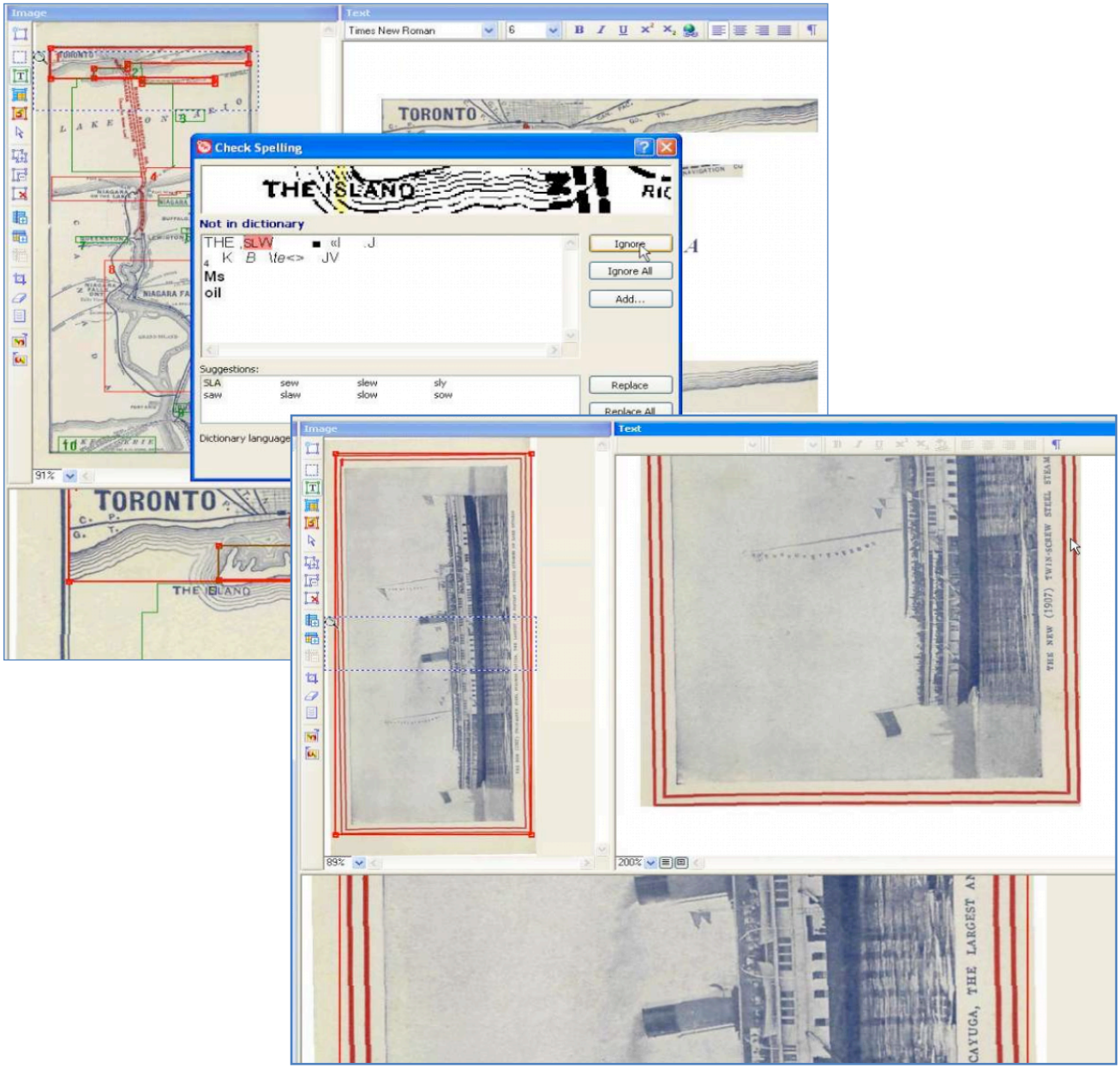
Note: Correcting OCR files

If the OCR translation of the files is not easily corrected (i.e. there is too much to correct for “full” text correctness), you can always correct keywords in the text file so that the object will be discovered through keyword searching. For incomplete correction files like this, please maintain the default “No” setting on the “Is this corrected OCR?” section on the Edit page screens.

Note: OCR processing of non-text content

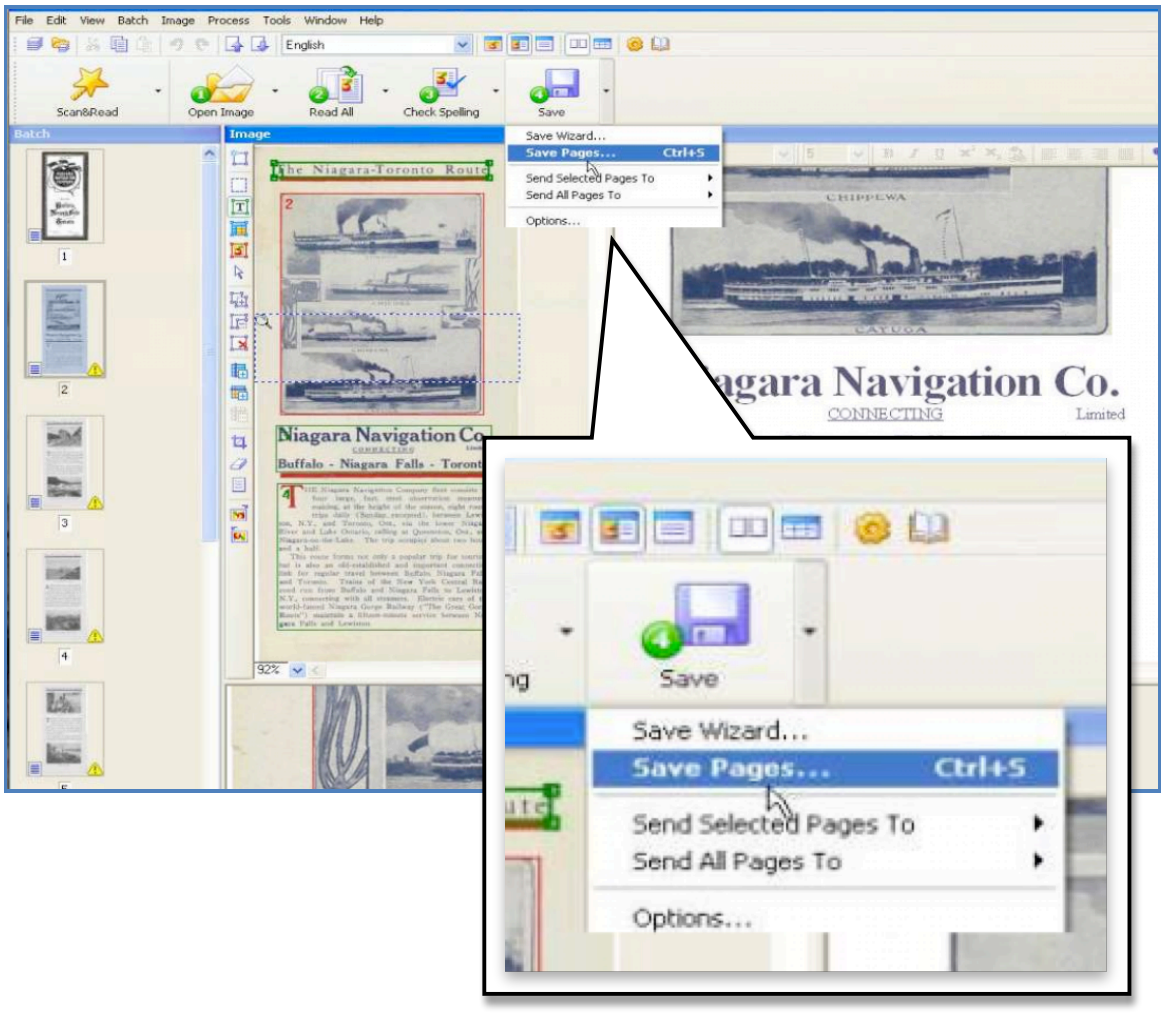
Unless you deselect a page from the thumbnails, ABBYY will try to read *everything* on the scans, including maps, images, landscape-oriented pages, etc. and this will return differing quality of text.

If you want the text on these pages to be full text searchable, you will have to correct or transcribe the words from these non-text pages manually.



Part 5.1.4 Save Pages

Once the pages have been “read”, click on Save and select Save Pages and this will activate the saving function for ALL the pages that were read (i.e. no .txt files will be saved for those pages that were deselected before the Read function was activated in the previous step)

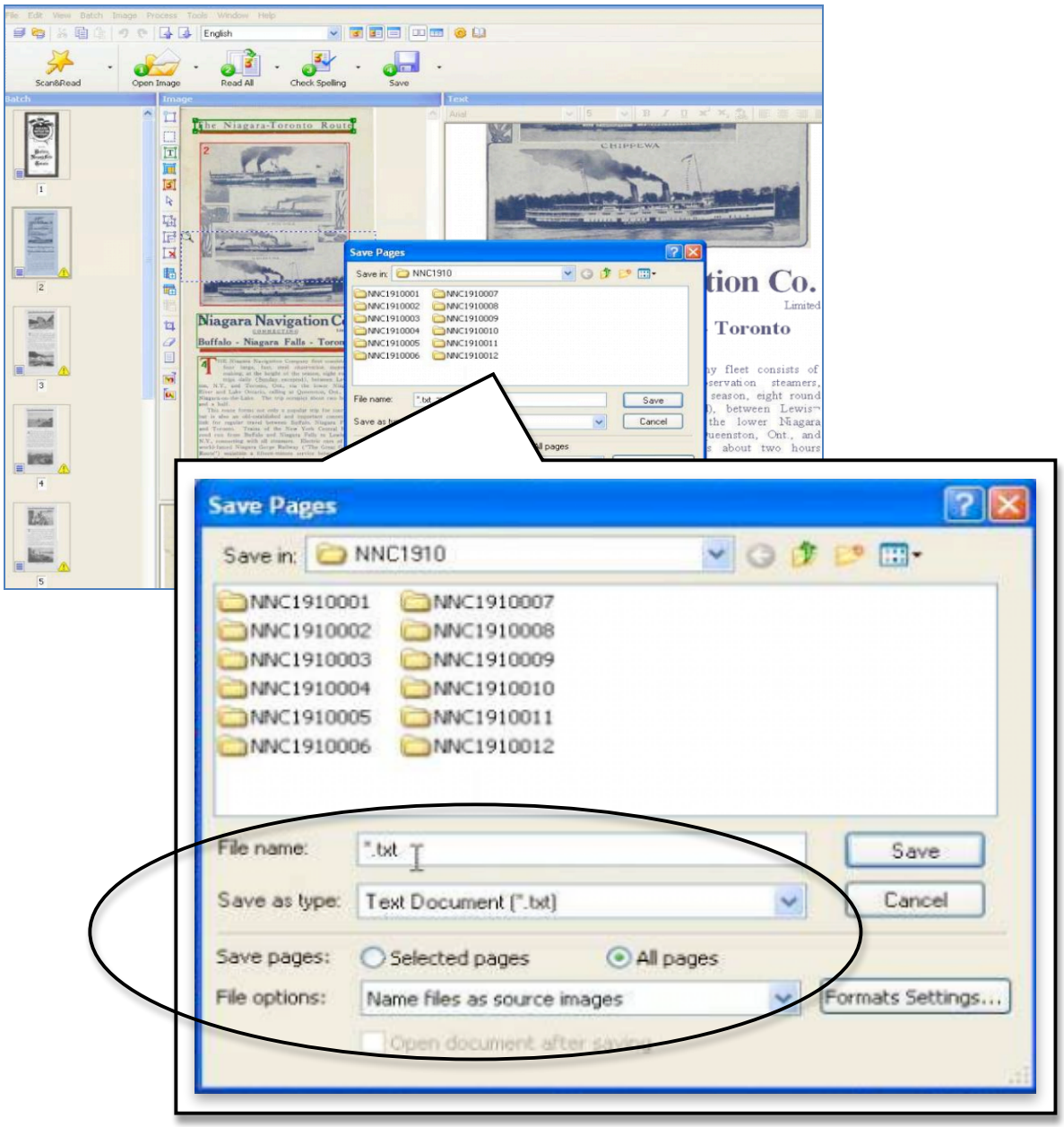


Save Files As…*

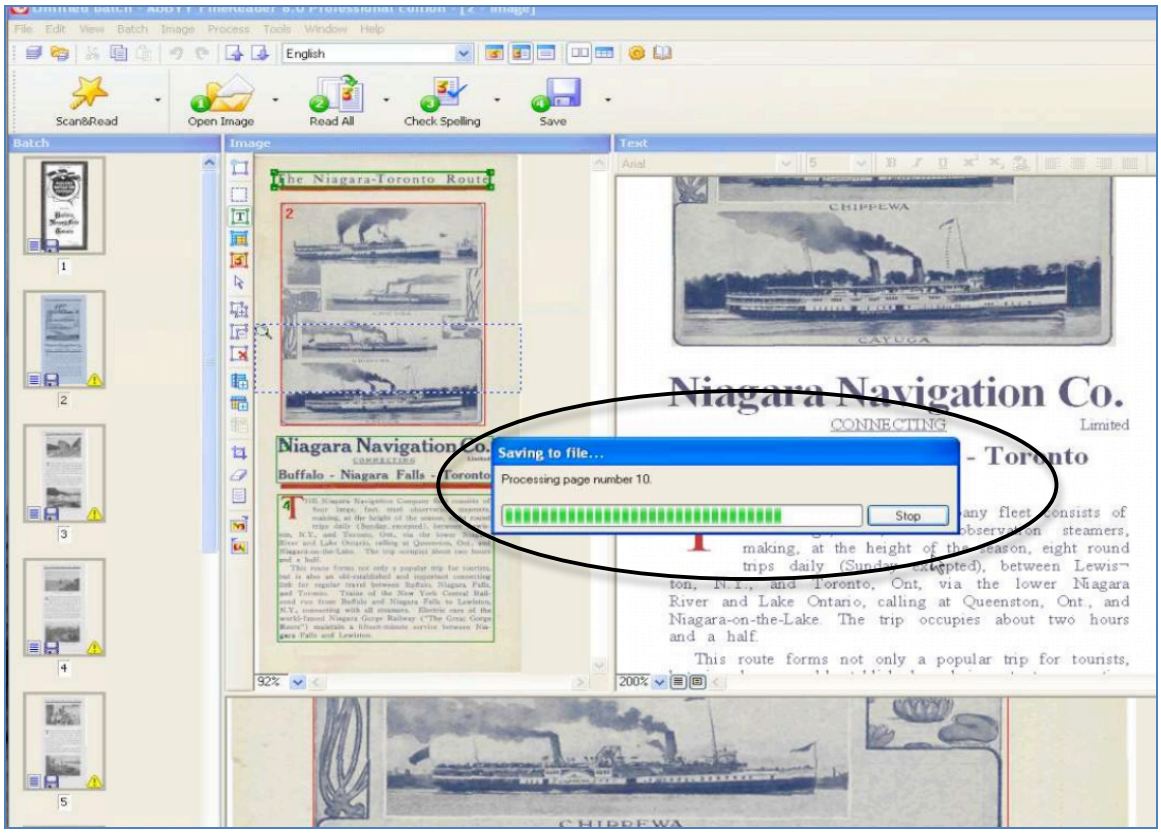
Confirm or select the following options:

- Save as type: Text Document (.txt)
- Save: “All Pages”
- File Options: “Name files as source images”

*The .txt files must be named exactly the same as the source image files to ensure that your full text will be correctly associated when you upload these files to the record using VITA.



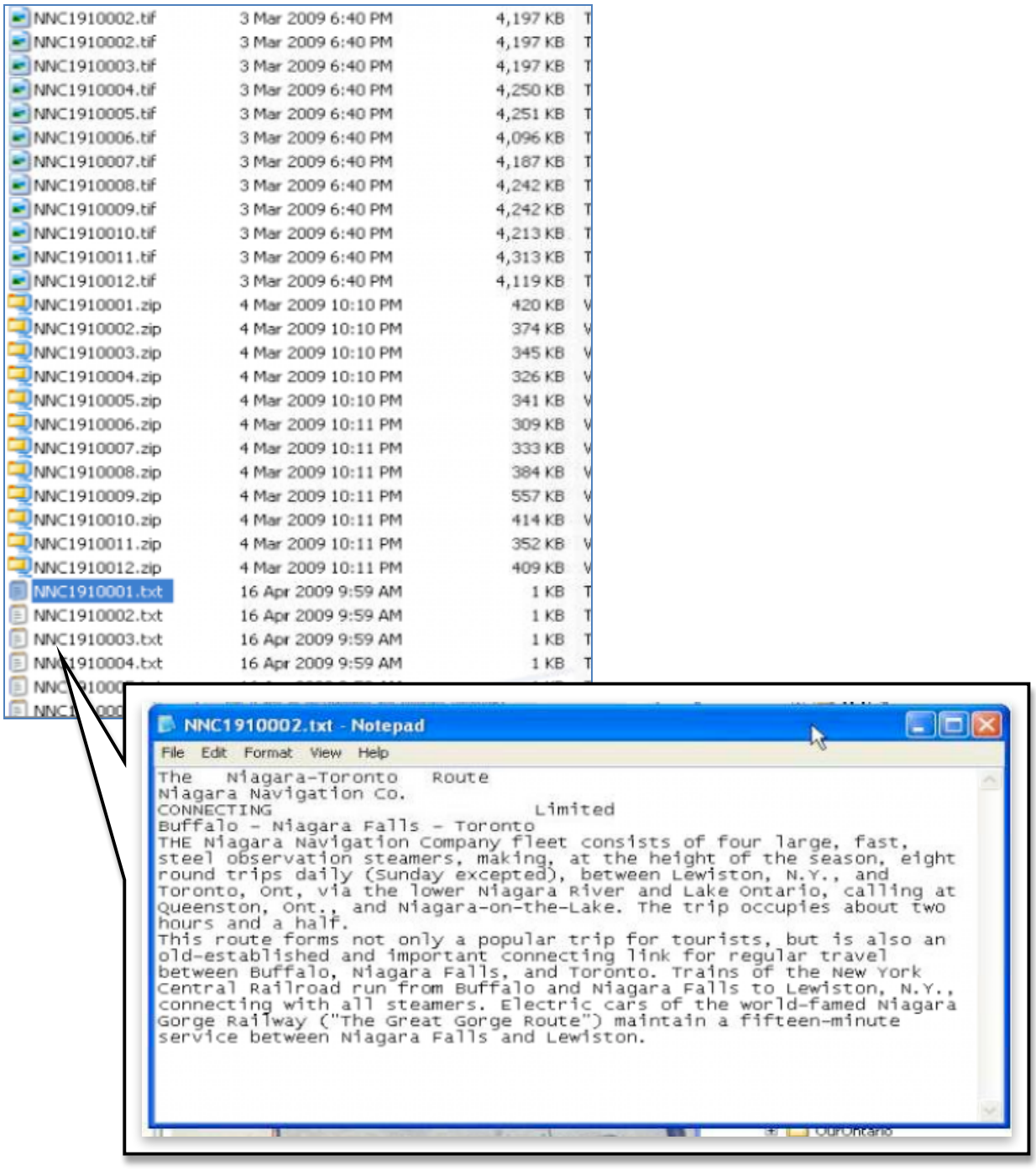
Wait while the application processes and saves the multiple files…



The .txt files are saved to the same folder as your source images. You should see a set of .txt files with exactly the same names as your .jpg or .tif files

These text files will open in Notepad or other basic text editing applications on your hard drive.

Note: Review the list of text files and eliminate any that are 0kb. The source page for that file likely has an image or was incomprehensible to the OCR software, but empty files will bog the upload process, so it is best to delete them beforehand and avoid frustration.



Part 5.1.5 Upload .txt files to VITA

Go to File/Tech data→Associate OCR/Full Text Files with this Record

Follow the steps outlined in Part 2 to upload single or multiple text files.

Make object full text searchable

Once the .txt files are associated with the record the OCR/Full file message will display in the File/Tech data screen for each page/file where an OCR file has been uploaded.

Click on the Details/Edit link to review and/or correct the full text and change the "Corrected OCR?" option to "Yes" if the text is *completely* correct. Click "Update/Edit Values"

In order for the text to be fully searchable, though, you must update the record to input that text for full text discovery on the user (public) side.

Go to the Descriptive or Administrative data screen and click “Update record”

When a user searches for a word that is discovered as part of the full text of your object, they will be directed to the page where the keyword is found.

Because the .txt files uploaded here are automatically assigned to the “Pages” category, the full text file will not display in the public side; only .txt file uploaded as category “Text” files can have their content displayed or suppressed in the public view. See Full Text Extraction for more information.

Unless you enter the full text from your text files into the Full text input box in the Descriptive data screen, this content is not displayed to the user (which hides any ugly OCR too).

For more details about the File Details/ Edit screen, see Part 2.4.