Using LEADTOOLS OCR to Enhance Google Drive Search

LEADTOOLS®
THE WORLD LEADER IN IMAGING SDKs

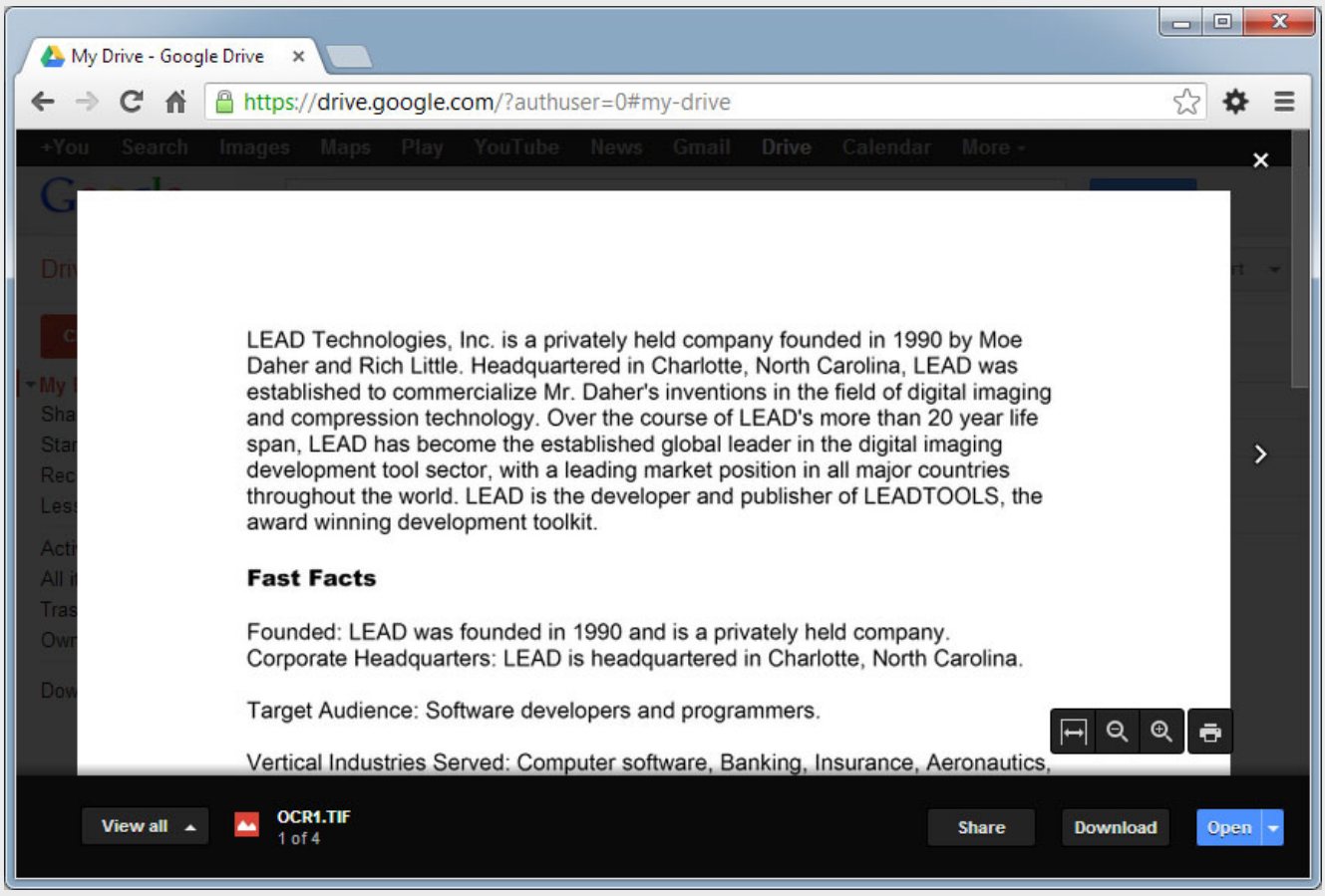
LEAD TECHNOLOGIES
INCORPORATED

1

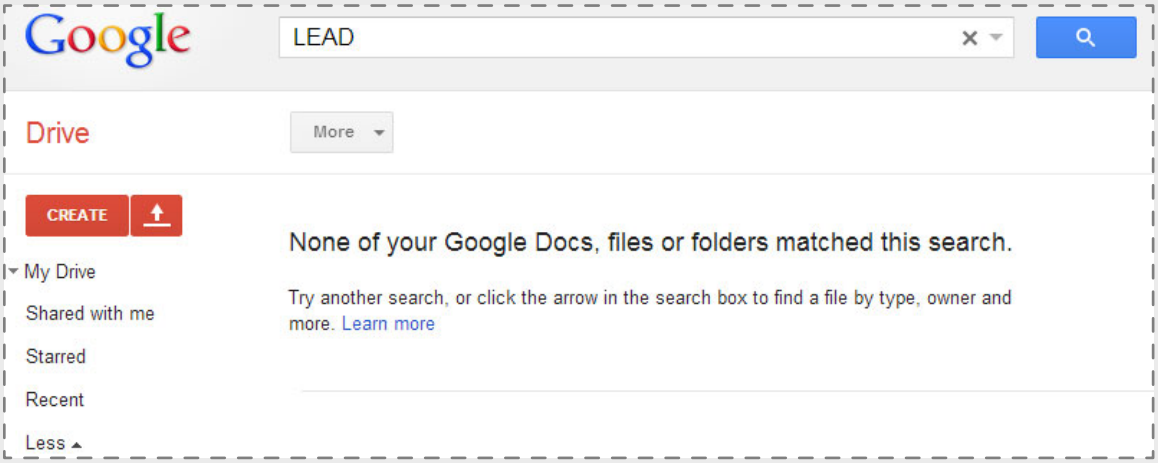
Introduction

Google Drive is a wonderful service for storing, organizing and sharing files such as documents, photos and videos. However, TIFF and other raster image file formats can get easily lost because Google Drive's search function can only do so much. With LEADTOOLS, developers can use its OCR SDK to extract the text and then add it to the `IndexableTextData` for each item. After this is completed, your raster image files can be searched in a similar manner to any text-based document like DOC or PDF.

For example, I have four ordinary TIFF files uploaded into Google Drive. Each of the four files are named OCR1 through OCR4, so only having the ability to search based on the file name isn't entirely helpful.



To the human eye, these images are nothing but text, but Google Drive only sees these images as raster data and returns nothing when I try to search for something internal to the scanned document.



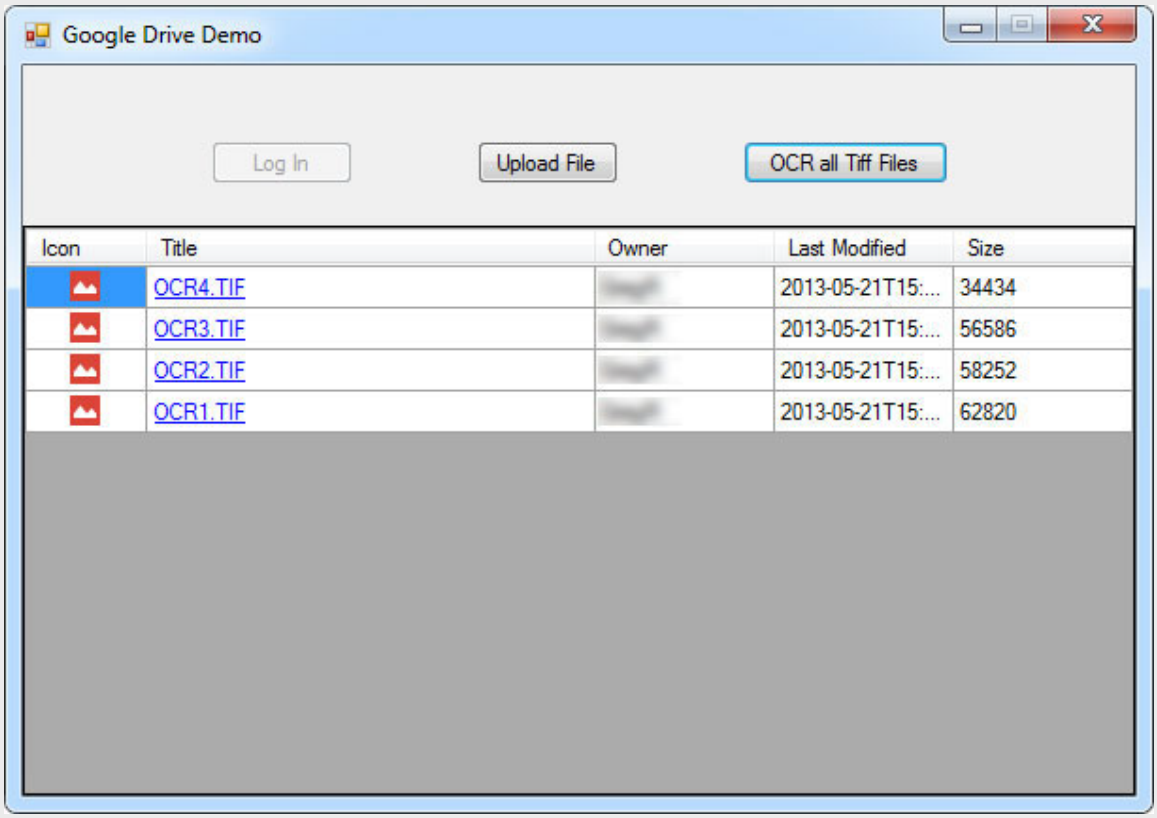
What would Google be without a way to search your files? Fortunately, Google Drive doesn't leave you hanging and uses the customizable "`IndexableTextData`" metadata of each document when it performs text search. In the example that follows, we show how to enable Google Drive to find these TIFF documents based on the text content without modifying the original image.

Connecting to Google Drive

The first step in this application is to enable the Google Drive API for our application to retrieve the `ClientID` and `ClientSecret`. We will need these properties later when using the Google Drive API for uploading and modifying the TIFFs. Lastly, we must download the Google Client Library to reference in our solution. For more detailed information on setting up a .NET application to interface with Google Drive, visit https://developers.google.com/drive/quickstart-cs.

In our application, we will open the User Authorization Uri in the `WebBrowser` control so the user can enter his Google username and password. After the user logs in, we can get the authorization code from the `WebBrowser` control's title. Now that the application is logged in and authorized to access Google Drive, we can search for all of the TIFF files in the account.

```
FileList fileList = googleDriveHelper.GetFilesList();

IEnumerable<File> tiffFilesEnumerable = fileList.Items.Where(
   file => file.MimeType == "image/tiff"
   && file.ExplicitlyTrashed != true
   && file.UserPermission.Role == "owner");

foreach (File file in tiffFilesEnumerable)v
{
   UpdateIndexableTextData(file);
}
```



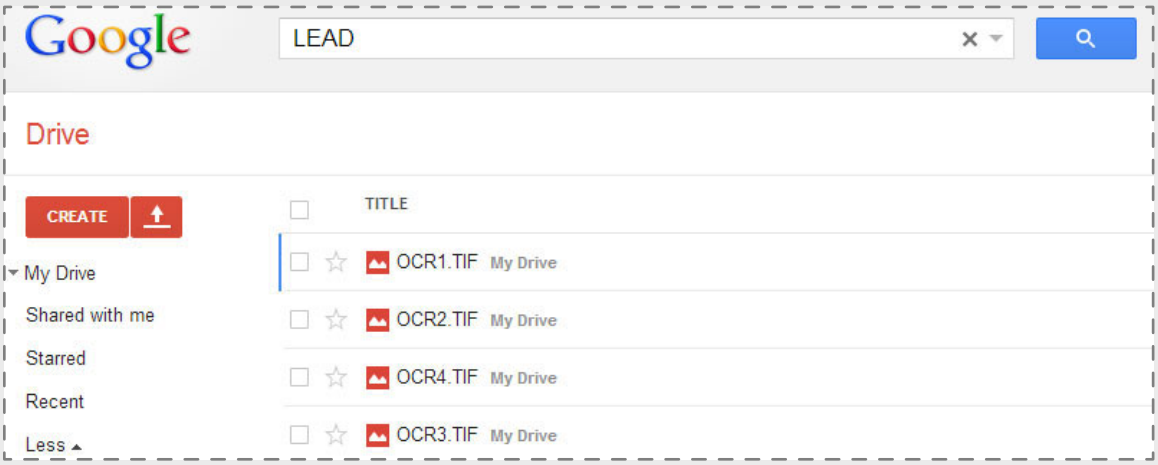
Using LEADTOOLS OCR

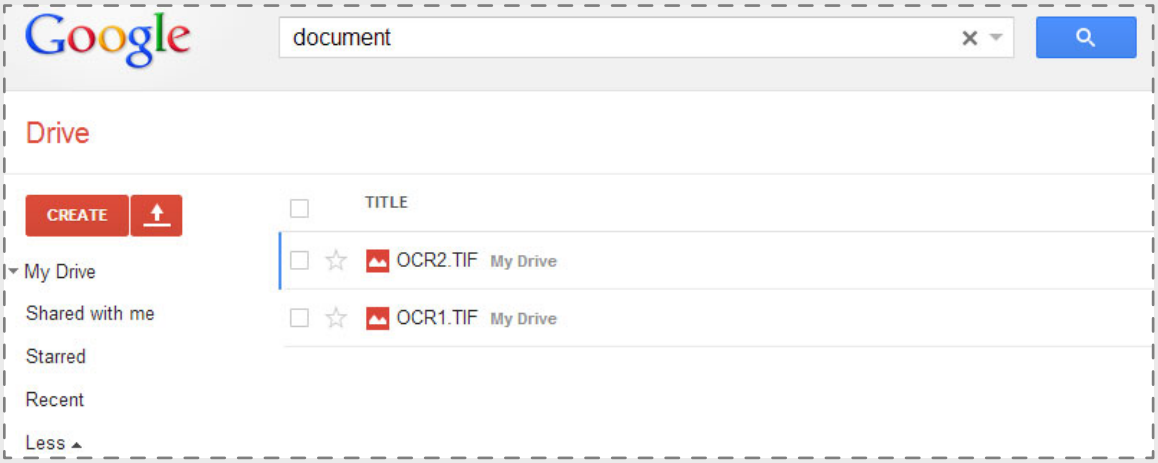
Finally, we can use the LEADTOOLS OCR engine to get the text for each TIFF file and all of the pages within it. After creating the `IOcrEngine` and `IOcrDocument`, the `RecognizeText` function will return a string value of all the text extracted from the page and then update the `IndexableTextData` metadata in Google Drive.

```
void UpdateIndexableTextData(File file)
{
   StringBuilder indexableText = new StringBuilder();

   // Get a .NET stream of the document
   using (System.IO.Stream stream = googleDriveHelper.GetFileAsStream(file))
   {
      // Create an instance of LEADTOOLS OCR engine
      using (IOcrEngine ocrEngine = OcrEngineManager.CreateEngine(
         OcrEngineType.Advantage, false))
      {
         // Start the engine using default parameters
         ocrEngine.Startup(null, null, null, null);

         // Get the number of pages in the document
         int pageCount;
         using (CodecsImageInfo imageInfo =
            ocrEngine.RasterCodecsInstance.GetInformation(stream, true))
```

```
            {
                pageCount = imageInfo.TotalPages;
            }

            // Create OCR Document
            using (IOcrDocument ocrDocument =
                ocrEngine.DocumentManager.CreateDocument())
            {
                // For each page in the document, recognize it
                for (int page = 1; page <= pageCount; page++)
                {
                    ocrDocument.Pages.AddPages(stream, page, page, null);

                    // Google Drive specific indexable text setup
                    indexableText.AppendFormat(
                        "<section attribute=\"Page{0}\">", page);
                    // Add the OCR text
                    indexableText.Append(
                        ocrDocument.Pages[0].RecognizeText(null));
                    indexableText.Append("</section>");

                    // Clear the document in prepartion for next page
                    ocrDocument.Pages.Clear();
                }
            }
        }
    }

    file.IndexableText = new File.IndexableTextData();

    file.IndexableText.Text = indexableText.ToString();

    googleDriveHelper.UpdateFileMetadata(file);
}
```

Now that we have processed all of the TIFF files in Google Drive, they can be searched by the text in the documents, even though they are technically raster images with no textual data.





Conclusion

This is just one of many real world solutions you can tackle with LEADTOOLS. You don't have to develop every concept and feature used within your product, but can enhance and leverage existing free cloud services like Google Drive to add incredible value for your customers and users. For more information on how LEAD Technologies can image-enable your application and boost your ROI, visit ***www.leadtools.com*** to download a free evaluation, or give us a call at +1-704-332-5532.

SALES: (704) 332-5532
SALES@LEADTOOLS.COM

SUPPORT: (704) 372-9681
SUPPORT@LEADTOOLS.COM

LEAD TECHNOLOGIES, INC.
1927 SOUTH TRYON STREET
SUITE 200
CHARLOTTE, NC 28203

About LEAD Technologies

With a rich history of over twenty years, LEAD has established itself as the world's leading provider of software development toolkits for document, medical, multimedia, raster and vector imaging. LEAD's flagship product, LEADTOOLS, holds the top position in every major country throughout the world and boasts a healthy, diverse customer base and strong list of corporate partners including some of the largest and most influential organizations from around the globe.

LEADTOOLS®
THE WORLD LEADER IN IMAGING SDKs

LEAD TECHNOLOGIES
INCORPORATED