Implementing a Standardized PDF/A Document Storage System with LEADTOOLS

LEADTOOLS®
THE WORLD LEADER IN IMAGING SDKs

LEAD TECHNOLOGIES INCORPORATED

1

Introduction

Electronic document archival has evolved far beyond the simple days of scanning a paper document and saving it as an image or PDF. Nowadays, many documents don't even start in physical form and could be one of many open or proprietary formats. Adding to the disparity caused by varying file formats is how and where files are stored. Many enterprises have their documents spread around numerous "data islands" including local computers, networked file shares and cloud services. Finally, the prevalence of mobile devices and tablets which may or may not support some formats further reinforce the need for standardized document archival.

Companies run on information, and as digital archives grow in both scale and diversity the ability to efficiently and accurately find data within them often fails to keep up. PDF/A is built for this purpose, but migrating all of your various file formats remains a challenge since raster image formats such as TIFF and JPEG have little to no searchable features beyond the file name. This white paper will explore how to take full advantage of PDF/A as your universal document storage format by using the state of the art technology within LEADTOOLS Document Imaging SDKs.

Creating a Searchable Document Archive with PDF/A

For years, PDF has been widely recognized and adopted as the best format for document archival, content management, record retention, risks management, litigation and discovery. This is especially true for the PDF/A sub-format which is specifically designed with archival and future-proofing in mind. PDF/A is completely self-contained and stores fonts, color management, annotations, images and more within the file itself. This ensures the document will stay true and not change its appearance for years on end while operating systems, devices, monitors and default fonts change all around it.

Normalizing your archive will yield many benefits in storage allocation, productivity and costs. The problem of being able to find and view your documents is drastically reduced since PDF is such a widely supported format. Making the choice to use PDF/A as your sole document archival format is certainly wise, but only solves a small part of the overall problem. Yet to be addressed are the issues of converting a divergent archive and ensuring that all further storage is done in a uniform fashion.

A handful of applications and scanners natively come with the ability to save as PDF, but can be unnecessary and cost prohibitive. In addition, documents can come from many sources both inside and outside your organization so at some level your documents must be processed and converted. Without a well designed and automated process, the benefits of a normalized archive are hard to fully realize. Many organizations therefore shy away from going fully digital due to the challenges involved in properly correcting and maintaining their newly envisioned document storage system. Therefore they feel trapped in knowing they need to change but do not know how to accomplish their goals in a holistic and cost effective manner.

Making it All Possible with LEADTOOLS Document Imaging SDKs

If all or part of this situation sounds familiar, look no further than LEADTOOLS. Its Document Imaging SDKs cover the gamut of imaging technology needed to make a universal PDF/A document archive a reality.

Full PDF and PDF/A File Format Support

LEADTOOLS provides full control over the PDF format including advanced capabilities such as extracting text, hyperlinks, bookmarks and metadata as well as updating, splitting and merging pages from existing PDF documents. With LEAD Technologies' decades of expertise in image compression, its PDF SDK also offers the industry's best performing and most diverse PDF compression options including JBIG, JPEG2000 and Mixed Raster Content. Also included are features often difficult to find in similar commercial SDKs, including reading, displaying, editing and writing native PDF annotations and markup that work seamlessly with Adobe Acrobat and other compliant PDF viewers.

Rather than being at the mercy of the PDF file format and the often exorbitant costs of PDF editing capabilities, LEADTOOLS will open up incredible opportunities for your archival system and keep all the decision making and customization in your court.

3

Optical Character Recognition (OCR) and Conversion

LEADTOOLS comfortably tackles the problem of migrating an existing archive with mixed file formats to a unified PDF/A archive. With the ability to load, save and convert over 150 raster, vector and document file formats, you can rest assured that you will have your bases covered.

Since not all formats are text-based and searchable, LEADTOOLS can use its fast and highly accurate Optical Character Recognition technology to convert those images to searchable PDF/A. The advanced OCR SDK in LEADTOOLS supports over forty languages and character sets including English, Spanish, French, German, Japanese, Chinese, Arabic and more, making it a reliable solution for the largest of enterprises running and providing services in multiple countries across the globe.

Most text-based PDF files also have smaller file sizes than the original raster image from which they were converted. Moreover, all of this can be done in as few as three lines of code.

```
IOcrEngine ocrEngine =
   OcrEngineManager.CreateEngine(OcrEngineType.Advantage, false);
ocrEngine.Startup(null, null, null, null);
ocrEngine.AutoRecognizeManager.Run(_strInputFile, _strOutputFile,
   DocumentFormat.Pdf, null, null);
```

Virtual Printing

If there is anything that the vast majority of applications have in common it's the ability to print. This is, after all, where the need for document archival started. Instead of printing documents to paper and then later using scanners and OCR to convert them back into a searchable digital medium, the LEADTOOLS Virtual Printer can get it done right from the start.

This approach not only handles the documents which you would normally print, but also allows you to archive many other sources of information including emails, faxes, website, social media and virtually any file format. As an added benefit, the vast majority of documents and materials you print are textual which means the resulting PDFs will already be searchable and require no special processing and are 100% accurate to the original document.

```
DocumentWriter _documentWriter;

public void _printer_EmfEvent(object sender, EmfEventArgs e)
{
   // Create a new document page and pass the EMF in e.Stream
   DocumentPage documentPage = DocumentPage.Empty;
   documentPage.EmfHandle = new Metafile(e.Stream).GetHenhmetafile();

   // Load EMF as raster for image over text
   e.Stream.Position = 0;
   documentPage.Image = _codec.Load(e.Stream);

   // Add the page
   _documentWriter.AddPage(documentPage);
}

public void _printer_JobEvent(object sender, JobEventArgs e)
{
   if (e.JobEventState == EventState.JobStart)
   {
      // Initialize DocumentWriter
      PdfDocumentOptions pdfOptions = new PdfDocumentOptions();
      pdfOptions.DocumentType = PdfDocumentType.PdfA;
      pdfOptions.FontEmbedMode = DocumentFontEmbedMode.Auto;
      pdfOptions.ImageOverText = true;

      _documentWriter = new DocumentWriter();
      _documentWriter.SetOptions(DocumentFormat.Pdf, pdfOptions);
      _documentWriter.BeginDocument(_pdfFileName, DocumentFormat.Pdf);
   }
   else if (e.JobEventState == EventState.JobEnd)
   {
      // Add fonts and end the document
      AddAndInstallFonts(e.JobID);
      _documentWriter.EndDocument();

      // Load PDF
      System.Diagnostics.Process.Start(_pdfFileName);
   }
}
```

Finally, LEADTOOLS Virtual Printers can also be configured to run on a server and made accessible over your company's LAN or the web with Internet Printing Protocol (IPP). This flexibility makes Virtual Printing an excellent solution for maintaining your archive into the future by providing a large funnel into which nearly any piece of information can be printed and then automatically archived through a central business workflow process.

5

HTML5 Zero Footprint Viewer

Just because you are saving your documents as PDF doesn't mean you can't benefit from a viewer. Though PDF is so widely adopted that few think about someone not being able to load it, plug-ins and viewing applications are still required in most situations. By using the HTML5 and JavaScript based viewer in LEADTOOLS, you can build a true cloud-based image viewing solution which requires no plug-ins or downloads. All of the heavy image processing and display is done on the client-side, yielding fast display times and a responsive user interface.

Conclusion

With LEADTOOLS, standardizing your document storage to PDF/A is no longer an arduous, complex and costly endeavor. Everything you need to convert your existing files, manage and normalize your PDFs, and create all-inclusive business workflows is included in programmer-friendly libraries for multiple platforms. You can rest easy knowing that all the information your company relies on for efficient and productive operation will be properly archived and readily accessible.

This is just one of many real world solutions you can tackle with LEADTOOLS. Its state of the art Document Imaging SDK is the most flexible and powerful product in its class, and LEADTOOLS offers an incredible value with its comprehensive family of toolkits for raster, document, medical and multimedia imaging. 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