End-to-End eDiscovery with LEADTOOLS Document Imaging

LEADTOOLS®
THE WORLD LEADER IN IMAGING SDKs

LEAD TECHNOLOGIES INCORPORATED

1

Introduction

When it comes to change, the desire for efficiency is surely at or near the top of the list of reasons. Some processes and industries are harder to change, especially those that have been around for a long time. Court systems in many countries are one of the oldest and most well established processes to ensure all-around fairness, even if it must sacrifice expediency. Thankfully, the legal industry has taken major strides towards adapting to the digital age with the evolution of eDiscovery and document imaging.

One major reason Electronic Discovery has become such a major part of the legal industry is the fact that so much data is natively generated and stored digitally. This electronically stored information (ESI) comes from a wide variety of communications (e.g. email, text messaging) and file formats. In addition to being a practical necessity, the benefits of eDiscovery are quickly expanding it into the preferred method of discovery. Rather than juggling two methods of discovery for different information sources, many legal counsels convert their traditional paper documents into ESI so all case information can follow the same process.

EDRM – Electronic Discovery Reference Model

Speaking of process, the Electronic Discovery Reference Model (EDRM) is a descriptive paradigm for how eDiscovery generally works in each case. Each stage in the EDRM is fluid and can be repeated or refined as the case evolves. The overarching goal and outcome is to take a huge amount of ESI and cull it down to what will actually be used in court.

Document Imaging technology, and in particular the technology offered in LEADTOOLS, has a role in nearly every stage of the EDRM. Some might be obvious, such as scanning paper documents into digital formats and using optical character recognition (OCR) to make the documents searchable. However, in a highly-competitive industry where time and funds can quickly get exhausted and turned into a lost case, having every available tool at one's disposal is vital to success in the courtroom. Before diving into the specific imaging technologies, the following overview of the EDRM will help enlighten how and where each technology fits in.

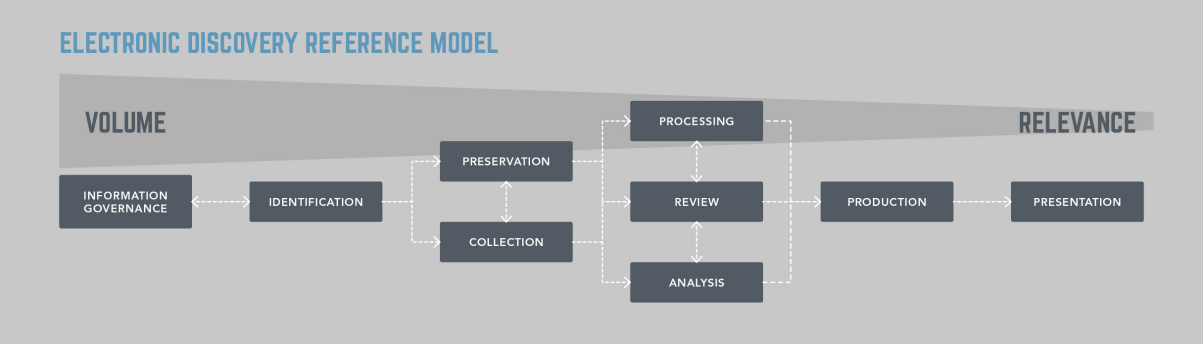

Figure 1: EDRM diagram (adapted from edrm.net)

Information Governance/Management

Information Governance, or Information Management ensures that proper information is saved and that it's stored and organized well.

Identification

This second phase in the EDRM is the first active step in the legal process, which throws a wide net to gather every bit of information that has any potential relevance to the case.

Preservation and Collection

After documents are identified, they must be preserved. Preservation applies a legal hold on ESI that requires the documents to not undergo any changes throughout the remainder of the case. Collection is the physical gathering of all ESI by a client into a transferable medium for sharing with their legal counsel.

Processing, Review, and Analysis

Removing duplicates and normalizing documents into decided-upon format(s) are the major tasks accomplished during Processing. Review takes a top-level look at the relevance of the ESI and Analysis is then a deep-dive look into the ESI to what is precisely relevant.

Production

Before they can appear in court, opposing sides must share their ESI with one another and decide which ESI is usable in the case.

Presentation

As implied by the name, the ESI is finally Presented in court.

3

LEADTOOLS at Work in eDiscovery Applications

LEADTOOLS Document Imaging SDKs have a wide gamut of imaging technologies perfectly suited in any eDiscovery application. The simplest of single-service specialists, end-to-end commercial ECMs, and everything in between will find everything needed to add world-class imaging technology to their applications.

Much the same way that the EDRM works as a general guide and process with steps that can be skipped or revisited, the imaging technology outlined below is not a mandatory set of features. For the most part, the order of these technologies follow the typical flow of use within an enterprise-level ECM but can be modified and reorganized to match the goals and creativity of any development team.

Scanning

Easily one of the most crucial elements of an ECM and eDiscovery application is the ability to digitize paper documents. Nothing is more efficient at getting high-quality digital replications of paper documents than scanning. Even if this is all that a company does to prepare for court, scanning provides massive savings in time and finances due to the simplified transportation and sharing of electronic documents. One USB stick can replace hundreds, if not thousands, of pounds of papers shipped and carried from one office to another and then to the courtroom.

LEADTOOLS includes high-level classes that make it very easy to acquire images from any scanner with a TWAIN driver or SANE backend. Consider the following snippet, which prompts the user to select a TWAIN source, then loads the acquired image into the viewer.

```
private void GetImageFromTwainSource()
{
   _twainSession.SelectSource(string.Empty);
   _twainSession.AcquirePage += new EventHandler<TwainAcquirePageEventArgs>(
      twainSession_AcquirePage);
   _twainSession.Acquire(TwainUserInterfaceFlags.Show);
}

private void twainSession_AcquirePage(object sender, TwainAcquirePageEventArgs e)
{
   rasterImageViewer1.Image = e.Image;
}
```

4

Document Cleanup and Preprocessing

Another hugely important feature when considering an imaging SDK is its ability to cleanup scanned images. There are two primary benefits to cleaning images, and each has huge trickle-down impacts on the entire eDiscovery process.

First, and probably the most obvious, is that the document itself is more readable. This is great for the human eye, but even better for the computer. Only a few pixels separate a lower-cased l, upper-cased L, and the number 1. The human eye can still read text with a strikethrough or a line caused by a crease in the paper, but even the best OCR engines will return gibberish.

Second, is storage space. Many compression algorithms accomplish their work by comparing neighboring pixels. This is especially true for black and white images that make up the majority of scanned documents. Performing image cleanup functions that remove dust speckles, hole punches, lines, borders, and the like have a profound impact on the length of runs and size of blocks comprised of a single color, allowing for very high compression ratios upwards of 92% of the dirty image's compressed size.

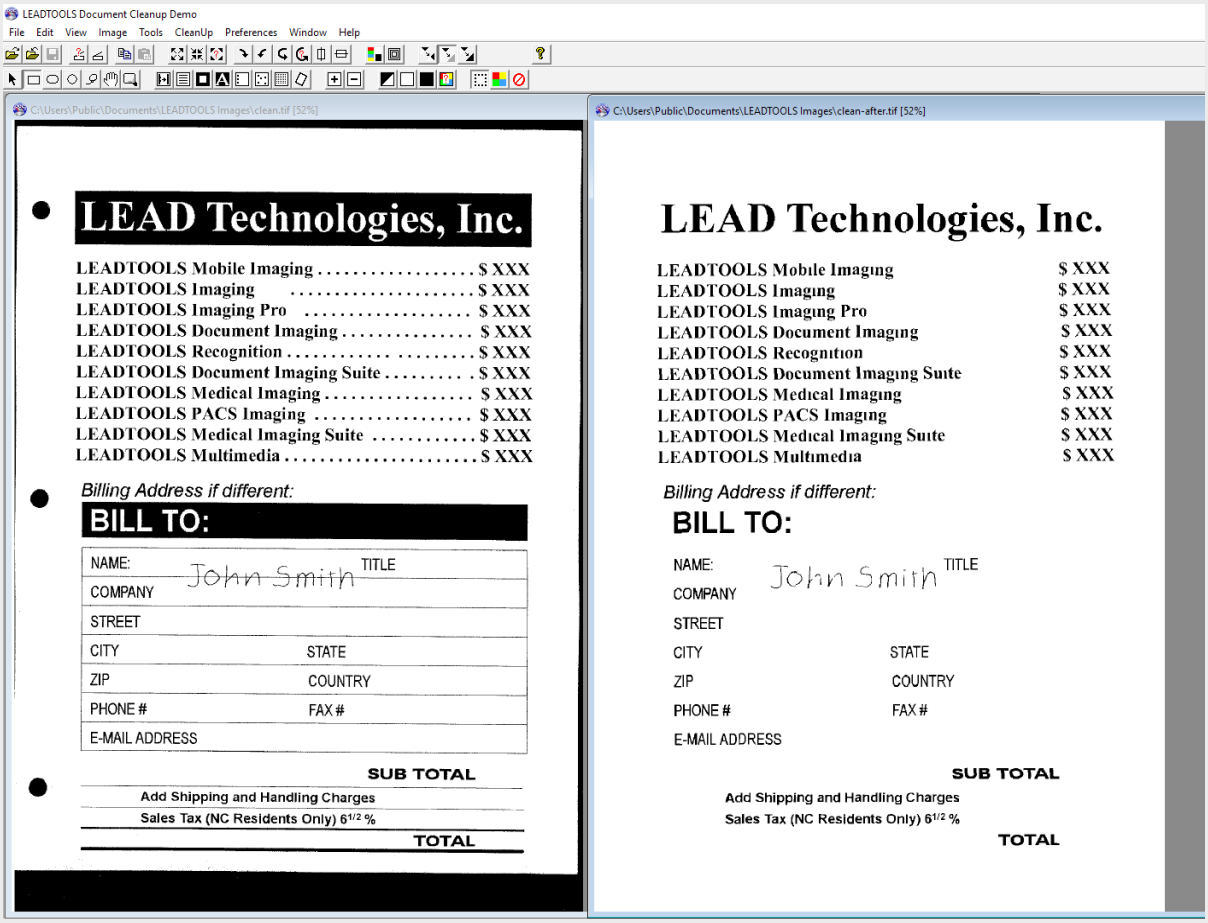

Figure 2: Dirty image cleaned with LEADTOOLS. Both are compressed with CCITT G4 but the clean image is only 12kb compared to 146kb

5

OCR

Optical Character Recognition is a must-have for any company wanting to step up their eDiscovery game. While scanning and cleaning images gets the job done and covers bases legally, OCR gives a competitive edge and is a major stepping stone into many other technologies.

Previously static images can come alive when converted to PDF, Microsoft Word, or any other number of searchable-text formats. Being able to search the contents of a digitally stored document is a monumental asset in eDiscovery, particularly during the Processing, Review, and Analysis phases. LEADTOOLS can create image-over-text PDFs, which are very useful in eDiscovery because the format retains the original image on top of a text layer, making it both searchable and virtually unchanged from the original. Even if TIFF is still the preferred file format within an ECM, OCR can infuse those static images with metadata that can be searched against or indexed for easier reference when needed.

LEADTOOLS makes OCR incredibly simple. Converting a source file on disk to a searchable PDF can be done with only three lines of code.

```
IOcrEngine ocrEngine = OcrEngineManager.CreateEngine(
   OcrEngineType.Advantage, false);
ocrEngine.Startup(null, null, null, null);
ocrEngine.AutoRecognizeManager.Run(@"C:\InputFile.tif", @"C:\OutputFile.pdf",
   DocumentFormat.Pdf, null, null);
```

Forms

Forms recognition is a specialized implementation of OCR that is finely tuned to extract specific pieces of information from documents rather than capturing the entire document. An organization may have a wide variety of forms ranging from tax forms submitted upon employment to invoices and bills sent to customers. By their very nature, forms have a lot of repeated information and perhaps the only relevant data is that which is filled in by the applicant, client, employee, etc. Therefore the most efficient and helpful means of archiving these forms within an ECM is to extract the data from those fields and either index them in a database or save them as metadata.

LEADTOOLS is able to handle both the classification and processing of forms. When classifying, LEADTOOLS will compare the incoming document against a repository of master templates (i.e. blank, unfilled versions of the form). Then once it finds a match, it will perform zonal OCR to extract the data from the filled out form.

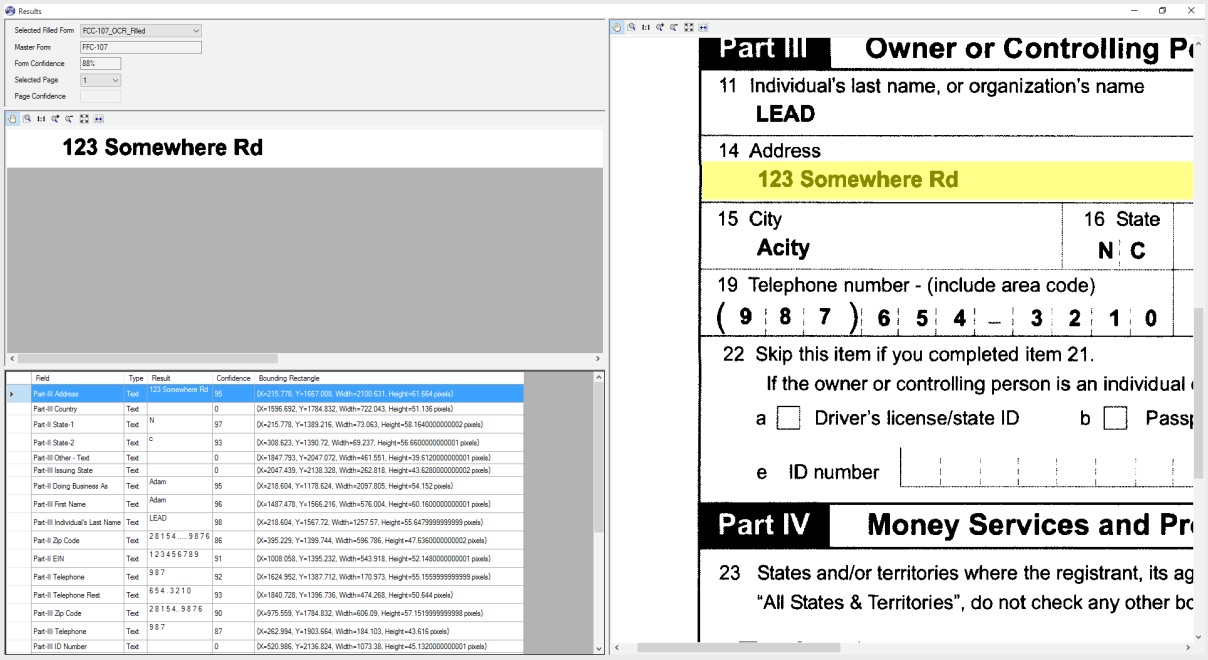

Figure 3: Displaying an extracted field using LEADTOOLS Forms Recognition

PDF

Adobe Portable Document Format is by far the most popular file format for storing documents, and with good reason. PDFs are easy to read, searchable, full of metadata, and can use annotations and markup. Its use is so commonplace that any concern over being able to load the file due to needing some kind of reader or plugin is largely moot. With all of that being said, other than the initial creation of a PDF through OCR mentioned earlier, why should an eDiscovery application developer use LEADTOOLS?

LEADTOOLS includes a full-blown PDF SDK that can accomplish far more than just creating a searchable PDF through OCR. Searchable text may be the biggest advantage of PDFs within the eDiscovery process, but unfortunately, that feature is typically confined to the viewer application. In a case with thousands of documents to analyze, opening each individual file is both tedious and prone to error. LEADTOOLS allows developers to parse the PDF document structure for the text, metadata, annotations, hyperlinks, and more, making it possible to search an entire folder or disk drive of PDF files for relevant ESI in a fraction of the time.

Document Converter and Viewer

Document conversion is a very common need during eDiscovery because each case is different and every client and litigator will come to various terms for the sharing of documents. An ECM may store everything as PDF, but the agreed upon format for production could be DOC.

7

LEADTOOLS includes a Document Converter library which can convert between a wide range of document formats including PDF, PDF/A, DOC/DOCX, XLS/XLSX, and PPT/PPTX *without the need for OCR*. This unique distinction is a huge benefit for eDiscovery because it satisfies the Preservation requirement while also simplifying Processing, Review, Analysis, and Production. The Document Converter can also take over 150 non-document formats like TIFF, JPEG, PNG, and BMP and use OCR to convert those images into PDF image-over-text, making it a perfect and programmer-friendly document normalization tool.

Closely related to the Document Converter, is the Document Viewer. On the surface, the Document Viewer, available for both .NET and JavaScript, implements many of the features that rich PDF viewing applications like Acrobat Professional contain: search and select text, follow bookmarks, add and edit annotations, and high-quality, vector-based zoom for smooth rendering at any scale factor. The Document Viewer really sets itself apart when combined with the Document Converter. Together, they offer on-the-fly document normalization that can view, search, and annotate over 150 document and image formats. Cloud-based storage and viewing is growing in acceptance across many courtrooms, and the LEADTOOLS JavaScript Document Viewer is a perfect fit that can view virtually any document in its native format with all of the same features as a text-based PDF.

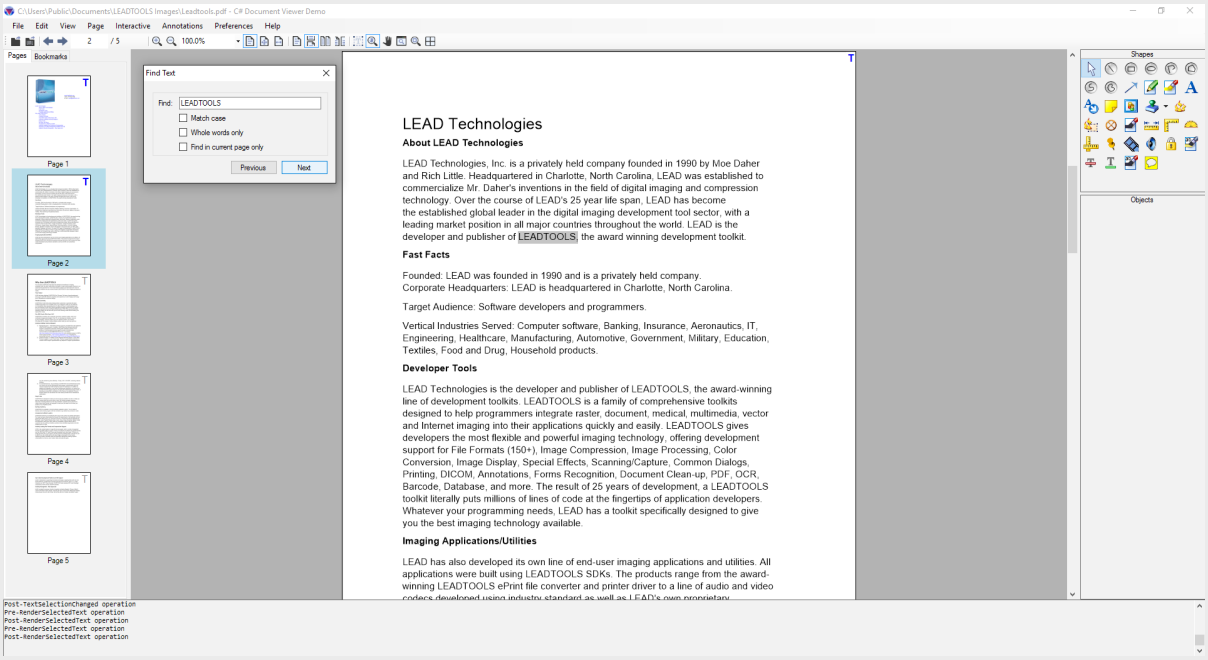

Figure 4: LEADTOOLS Document Viewer in .NET application showing page thumbnails, annotations, and searchable text.

Annotations and Markup

Annotations and Markup can show up in many stages of the EDRM. Annotations exist in a layer on top of the image or document and can infuse the document with additional information and encourage collaboration and communication. Sticky notes, arrows, and highlights can draw attention to important parts of the document.

Probably the most important annotations for the legal industry are bates stamps, redactions, and encrypt. Bates stamps have a long history in the courtroom of sequentially stamping paper documents for identification purposes. Digital ESI is no different and LEADTOOLS annotations make it easy to overlay consecutive numbers on documents as they get imported into an ECM or transferred into a new storage disk for Production. Redaction uses black rectangles to cover up sensitive information that should not be divulged in court. Digital redaction with annotations is beneficial since it doesn't require any changes be made to the original ESI. Redaction annotations can even be undone with passwords, providing different amounts of information based on a user's access tier. The encrypt object works similarly to redaction, but scrambles the image underneath of it rather than completely hiding it.

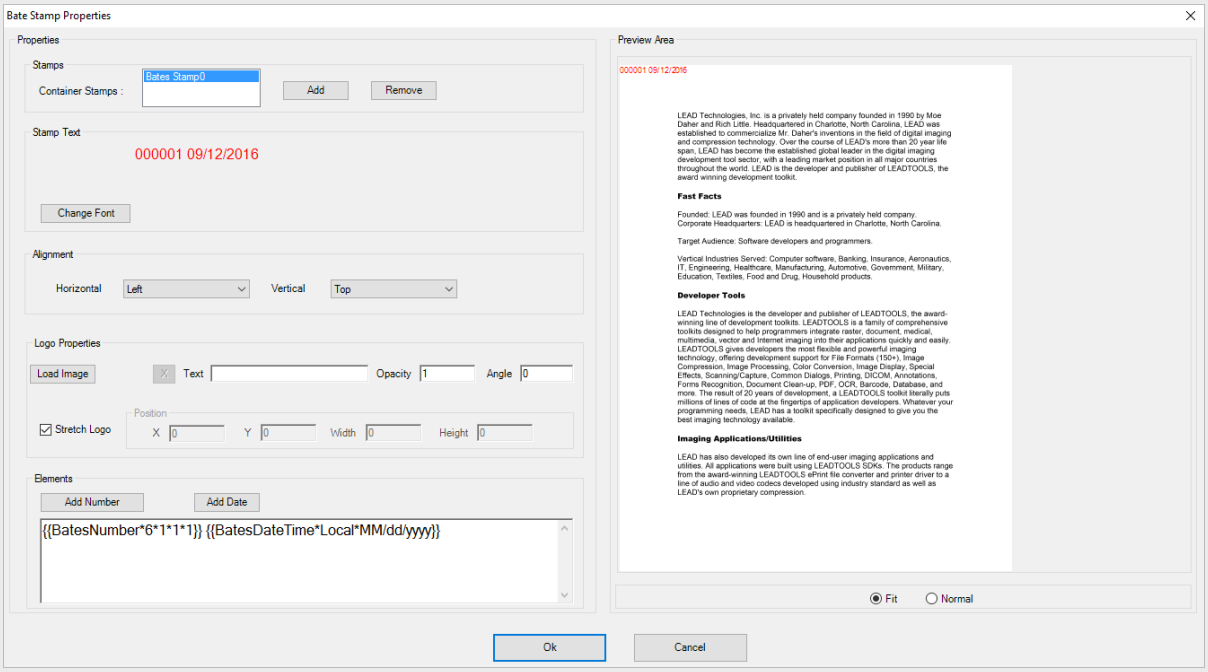

Figure 5: LEADTOOLS Bates Stamp Properties Dialog

Virtual Printer

There are some instances where ESI is difficult to process because it is saved in a format only usable in a legacy application or custom-developed system. Thankfully, the vast majority of applications have some form of printing functionality, but many legal teams don't use this to their advantage. Virtual Printing, also known as print-to-file drivers, is an oft-ignored solution that helps prevent the hassle of printing to paper and then using a scanner to import it into an acceptable electronic format.

The LEADTOOLS Virtual Printer offers a unique catch-all solution by capturing the print job from any application and converting it to LEADTOOLS-supported formats, including searchable PDF, DOC, and TIFF. Adding annotations is also supported within the Virtual Printer, allowing developers to create workflows that can tackle several goals at once by both normalizing documents into appropriate ESI and bates stamping them along the way.

Conclusion

eDiscovery is a huge market for software developers, full of opportunities at any point of the legal process. Businesses both large and small have various requirements, from preventative and organizational measures, finding and analyzing ESI, creating ESI from paper, all the way to presenting their case in the courtroom. With excellent support for scanning, document cleanup, OCR, forms recognition, PDF, document conversion and viewing, annotations, and virtual printing, LEADTOOLS Document Imaging SDKs can handle every requirement software developers in the legal industry are regularly asked to implement.

LEADTOOLS offers an incredible value with its comprehensive family of toolkits for document, medical, multimedia, and raster imaging. For more information on how LEAD Technologies can image-enable applications and boost ROI, visit www.leadtools.com to download a free evaluation, or give us a call at +1-704-332-5532.

Sales: +1(704) 332-5532
sales@leadtools.com

Support: +1(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 25 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