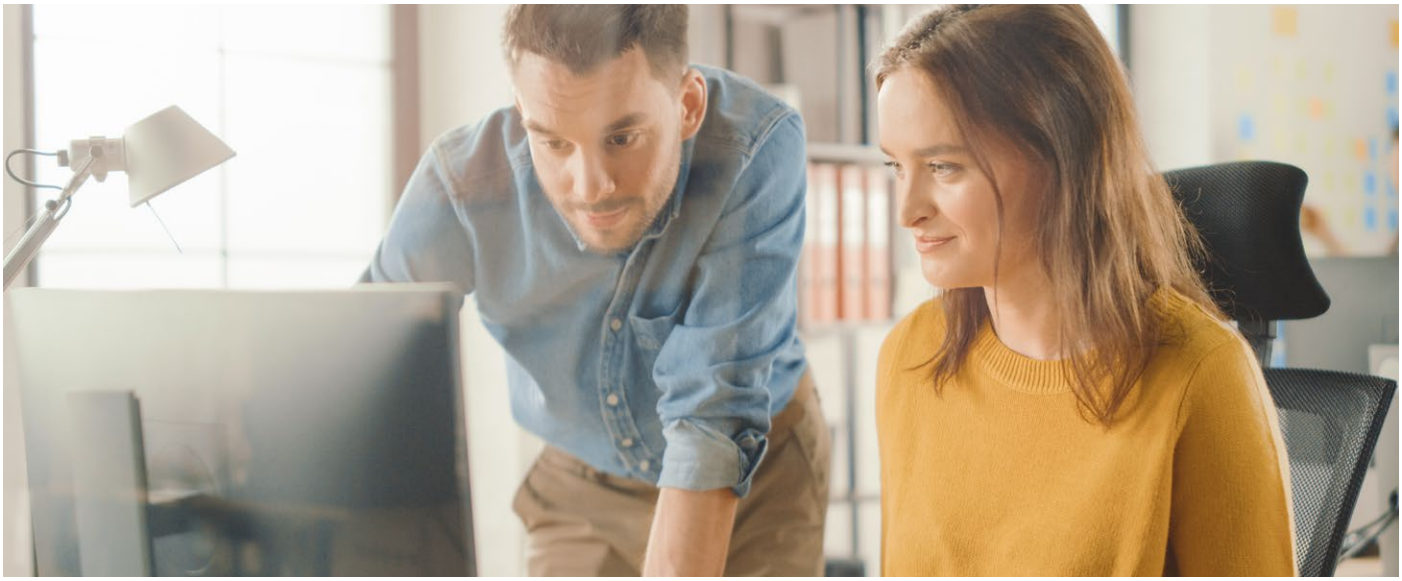
OpenText File Content Extraction

Help customers get more out of their data with accurate file format identification, content decryption, text extraction, subfile processing, nonnative rendering, and structured export



Benefits

- **Reduce time to market,** engineering risk, development costs, and maintenance
- **Don't miss a thing,** with the richest text extraction technology available
- **Improve usability** by providing high-fidelity HTML renderings

Established corporations need support for a variety of historic and modern office applications to ensure data coverage. Startups and newer companies often rely on newer cloud native formats. If you target governments, good support for open formats can be a deal maker. Some industries have a particular class of file formats that are disproportionately important but not well supported elsewhere, such as the compliance market's requirement for historic office files.

OpenText™ File Content Extraction is a mature and professionally maintained OEM embeddable OpenText™ Knowledge Discovery SDK for file format identification, content extraction, and file transformation used by leading edge software developers and service providers to add competitive differentiation and to significantly reduce the business risks associated with managing large volumes of human originated information.

Reduce time to market

OpenText File Content Extraction is designed with OEM in mind. Unique applications have unique demands of the technology they embed. OpenText understands this, providing a range of options to suit multiple deployment environments: endpoint, server/cloud, and hybrid models, with support for Windows, Linux and macOS, on both Intel and ARM architectures, with native APIs for a variety of programming languages. Easily integrated libraries and reference code mean quick integration with new or existing applications.

Kainos Evolve – The UK's leading EMR company used OpenText Knowledge Discovery to create its flagship product Evolve EMR, which supports vital analysis of healthcare data.

- Eliminates paper records in line with UK government directives.
- Enables effective searches to increase staff efficiency, and delivers holistic intelligence to improve strategic decision-making
- Provides integrated solutions that speed up development and time-to-market

Read full story >

OpenText File Content Extraction supports both file-based and stream-based I/O for best fit with any application architecture.

Secure by design, OpenText File Content Extraction minimizes risk through techniques such as attack surface reduction for reduced threat impact, component isolation for quick third-party vulnerability mitigation, process privilege reduction, and countermeasures for specific format-based attacks, such as Zip Slip. A thread-safe, in-process or out-of-process model for threat containment, improves application stability in the face of any input data. We understand that one of the biggest security threats to your product is your supply chain. Flexible licensing ensures that we can align with your business model.

Don't miss a thing

With continuous development for more than 25 years, OpenText File Content Extraction has an extensive catalogue of supported formats, kept up to date by our professional team to address the relentlessly increasing number of relevant document formats—everything from legacy formats that current software cannot read, to formats from new applications and software updates, as well as nascent cloud-specific formats.

Accurate file type identification is crucial for determining the correct downstream processing. File formats vary massively in terms of the amount and accessibility of usable content. It is important to correctly identify even file types that do not contain useful content, to make an accurate risk assessment for use cases such as DLP and eDiscovery.

Improve usability by providing high-fidelity HTML renderings

Extract structure from document content, creating well-formed XML which is validated against a predefined Document Type Definition (DTD). OpenText File Content Extraction applies an XML vocabulary to the data structures in a document so that downstream applications can access content in context. It returns structure such as headers/footers, footnotes, endnotes, bookmarks, headings, sheet names, and structured table data.

Preview documents in high-fidelity HTML. Incorporating this technology into your web-based applications enables your end users to view a document even if they do not have the appropriate plugin or native application.

With HTML Export, you control the content, structure, and format of the HTML output using easily customized templates or the flexible and robust APIs. Choose between web friendly dynamically flowed text for the best understanding of a document's content or a fixed width rendering, mimicking printed output.

Break content into manageable chunks for a faster load time and lower browser memory footprint. Add structure, such as highlighting using custom markup, and automatically generate a navigable table of contents based on document properties, such as font size or style. Apply Cascading Style Sheets (CSS) to improve output fidelity and align look and feel for a quick and easy read.

File format detection

Reduce the risk of misprocessing crucial information or wasting valuable CPU time on irrelevant files by quickly and accurately identifying file type. Instead of relying exclusively on falsifiable file name extensions or short magic numbers, OpenText File Content Extraction forensically examines each file, focusing on the most differentiating characteristics first and going as deep as needed to resolve ambiguity, resulting in faster answers and a lower error rate.

OpenText File Content Extraction goes beyond MIME type, clearly identifying files with non-existent or ambiguous MIME types (e.g., application/octet-

Censornet – Censornet's autonomous, integrated cloud security empowers mid-market organizations with enterprise-grade protection. It offers email, web, cloud application security (CASB), and identity solutions for robust cyber protection.

- Enhances data protection
- Ensures regulatory compliance for comprehensive cybersecurity
- Provides competitive advantage with easy market expansion opportunities

Read full story >

stream), adding detail such as encryption status, format classification, and format version, allowing you to make your downstream routing and processing decisions with precision. Supported document classes include analytics, animation, CAD, database, desktop publishing, encapsulation, executable, font, GIS, library, movie, object module, outline, presentation graphics, raster image, schedule, scientific, sound, source code (including language identification), spreadsheet, vector graphics, and word processing.

Rights management

Identify rights management protected files from Microsoft, Seclore and SmartCipher. Inspect MSIP (Microsoft Purview Information Protection) labels, even from encrypted files, to correctly determine risk. Decrypt files, credentials required, that have been protected by Microsoft Azure Rights Management (RMS), part of Microsoft Information Protection and associated technology, allowing your workflow to operate transparently on the original, unencrypted content.

Metadata access

Quickly access file metadata such as XMP, XrML, IPTC, EXIF, Boldon-James classification, and format specific fields. File Content Extraction combines and normalizes common fields for easier downstream consumption.

Character set conversion

Prepare for downstream processes, which usually expect UTF8 input. OpenText File Content Extraction automatically determines the character set used within a document, even if this is not specified in the metadata, and converts this to UTF8 or another encoding of your choice. With correct identification and conversion, value is maintained.

Text extraction

Extract plain text content by removing format scaffolding and other noise at speed. Go deep into a variety of document formats, extracting body text and other visible components (such as headers, footers, footnotes, endnotes, captions and table components), with the option to include hidden text (such as section names, notes, tracked changes, explicitly hidden elements, accessibility layers and configurable placeholder text), as well as cached, orphaned, unused and deleted text.

Subfile extraction

Dig into formats that commonly embed further content, from the obvious archive formats and email stores to more surprising container formats, such as PDF and its variants, and all of the big three office type documents: word processing, spreadsheet and presentation graphics. Through OCR, directly access the textual content of scanned documents, photographed receipts, and raster images containing text, as part of the processing pipeline.

Conclusion

Gain competitive differentiation and significantly reduce the business risks associated with managing large volumes of human originated information. OpenText™ File Content Extraction provides a mature and professionally maintained OEM embeddable SDK that identifies file formats, extracts content and provides file transformation that ensures nothing is missed and time to market is reduced.

	APIs					Platforms						
	C	C+	Python[1]	Java	NET[2]	Windows/ x86_32	Windows/ x86_64	Windows/ ARM[3]	Linux/ x86_64	Linux/ ARM	macOS/ x86_64	macOS/ ARM
File Format Detection	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rights Management[4]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Meta Data Access[5]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Character Set Conversion	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Text Extraction	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Subtle Extraction	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
OCR	✓	✓	✓	✓	✓		✓		✓	✓		✓
Structured Export	✓			✓		✓	✓		✓	✓	✓	✓
HTML Export	✓			✓		✓	✓		✓	✓	✓	✓
PDF Export	✓						✓		✓			

✓ Available

1 Python availabe only x86_64 platforms and macOS/ARM

2 .NET available only on windows platform.

3 Windows/ARM supported only for C API

4. Decryption (separately licensable) available only in C and Java APIsk on Windows/x86_64 and Linux/x86_64 platforms

5 .NET metadata access excludes normalization

Resources

Request a demo >

Learn more >

What's new >

File Content Extraction deployment options:

Extend your team

- On-premises software, managed by your organization or OpenText

About OpenText

OpenText, The Information Company, enables organizations to gain insight through market leading information management solutions, on-premises or in the cloud. For more information about OpenText (NASDAQ: OTEX, TSX: OTEX) visit: opentext.com.

opentext™