OpenText File Content Extraction

辦識、擷取及轉換檔案內容

優點

- 縮短上市時程、降低工程風險及內部開發成本，並減少持續維護作業。
- 讓開發人員無需花時間隨時掌握不斷變化的檔案格式環境，目前支援 1,950 種格式。
- 降低不當處理重要資訊的風險。
- 避免浪費寶貴的 CPU 時間在不相關的檔案上。
- 採用市面上最豐富的文字擷取技術，不會遺漏任何資訊。
- 使用 OCR 提升所有文件來源的可見度。
- 以子文件精細度處理資料管道以降低延遲。
- 提供高精確度的 HTML 轉譯功能以提升易用性。
- 利用 KeyView 的安全程序外功能來提高穩定性。
- 支援權限管理功能 (例如 Microsoft Purview Information Protection)，協助客戶因應新興安全資料共享趨勢。

提供準確的檔案格式辨識、內容解密、文字擷取、子檔案處理、非原生轉譯及結構化輸出，並支援所有主要用戶端與伺服器端平台上 1,950 種格式，讓您的客戶能夠更充分地利用資料。

產品焦點

OpenText File Content Extraction 是經過專業維護的成熟 OEM 嵌入式軟體開發套件，用於辨識檔案格式、擷取內容及轉換檔案，頂尖軟體開發人員和服務供應商均使用此產品來增加競爭優勢，並大幅降低管理源自人類之大量資訊的相關業務風險。

加速實現價值

OpenText File Content Extraction 是將深入內容可見度整合至您的服務或應用程式中的低風險方法，不僅快速、可靠，而且不需要持續開發。

我們提供現成可用的軟體開發套件，並提供範例程式碼來加速產品上市時程，讓您的工程組織有時間投入企業的核心價值主張。

我們會隨時掌握充滿新格式及不斷演進之格式的複雜環境，讓您可以省下時間心力。

涵蓋範圍更廣

請想想您銷售的目標產業和地區。

創立已久的企業需要支援各式各樣過去和現今使用的辦公室應用程式，以確保能提供適當的資料涵蓋範圍。新創公司和較新的公司通常會仰賴較新的雲端原生格式。如果您的目標對象是政府 (尤其是在歐洲或南美洲)，針對開放格式提供完善支援有利於促成交易。有些產業具有特定的檔案格式類別，其重要性極高但通常無法在其他地方獲得妥善支援，例如法規遵循市場對於過去辦公室檔案的要求。

在格式支援方面，地理多元性扮演了重要的角色。即使是在辦公室應用程式等「簡單」的領域，不同地區事實上也有不同的應用程式與檔案格式。舉例來說，Hangul 辦公室套裝軟體 (本身具有一組不斷演進的格式) 因為提供完善的韓文支援，而在南韓廣為普及。在中國，國內的 WPS 辦公室套裝軟體 (前身為 Kingsoft) 則與 Microsoft Office 互相競爭。

涵蓋範圍廣度也適用於語言支援：PDF 文件若包含由右至左的語言或含有偏移字元的語言 (例如阿拉伯文和希伯來文的母音)，可能會難以擷取正確的文字。

您計畫如何擴展業務？您接下來鎖定的目標是何處？您的解決方案是否已準備好應對您在該市場中會遇到的獨特格式？

OpenText File Content Extraction 已持續開發超過 25 年，擁有豐富的支援格式目錄，並由我們的專業團隊隨時維持在最新狀態，以因應不斷增加的相關文件格式數，從目前軟體無法讀取的舊格式，到新應用程式和軟體更新的格式，以及新興的雲端特定格式，全部都包含在內。

更深入的涵蓋範圍

準確的檔案類型辨識功能對於判斷正確的下游處理方式至關重要。檔案格式在可用內容的數量和可存取性方面有極大的差異。正確辨識檔案類型對於準確評估 DLP 和 eDiscovery 等使用案例的風險相當重要，即使是不含實用內容的檔案類型亦然。

即使沒有實用的文字內容，系統仍會擷取中繼資料。 這是檔案上的主要資訊來源，對於驅動正確的行為至關重要。

文字內容通常是您在原生應用程式中開啟檔案時會看到的內容。這不僅包括本文，還有儲存在其他位置的可見文字，例如標題、浮動區塊、圖表字幕、影像標題、註腳和註解，以及其他頁面項目。標題是否寫出「公司機密」？

若辨識為表格式內容，則會對下游處理特別實用。表格標題可提供有關儲存格內容真實意義的重要線索。例如有了「社會安全號碼」這個標題，就能讓系統理解未標示的字串，避免出現難以管理的大量誤判結果。

隱藏內容是許多格式的重要部分，其中包含檢視畫面時可見內容之外的寶貴內容，例如明確隱藏的區段、追蹤變更 (包括刪除的文字)、快取內容、指令碼/巨集、孤立物件、協助工具內容，以及根據檢視器功能呈現的多種動態轉譯表現方式 (例如使用 DOCX 的 Alternate Content / Choice / Fallback 功能)。

容器檔案會內嵌其他檔案，也就是子檔案。若要完整檢視，則必須拆解這些檔案。雖然最顯而易見的容器檔案是 ZIP 或 RAR 等歸檔，但有許多常見的非歸檔檔案類型也會包含子檔案，例如 PDF (一般和文件夾)，Microsoft Office (OLE 和 XML 格式)，以及 PST 等信箱格式。容器格式與壓縮編解碼器均有多種標準，某些容器格式支援額外的元件，例如替代的資料流，以及根據用於擷取的平台而產生的不同內容。

大多數的檔案格式不會公開記載，即使是公開記載的格式，使用其公布的規格作為粗略指引可能會比作為實際資訊更為實用。實作各有不同且包含錯誤，而要真正瞭解格式在實際環境中的外觀細微差異，必然需要進行一定程度的逆向工程。

許多地方都能找到使用者建立的內容，而其中有許多內容並不顯而易見。您遺漏了什麼？OpenText File Content Extraction 是高可見度選項，可揭露其他解決方案無法觸及的內容。

效能更高

盡可能提高輸送量、將延遲降至最低、降低 CPU 成本、縮小安裝大小、最佳化記憶體佔用空間，效能包含眾多層面。OpenText File Content Extraction 是為了成為您文件管道中重要的一環而設計，絕不會成為瓶頸。

快如閃電的準確格式偵測功能會執行所有必要工作 (但僅止於所需作業) 以達成高信心度的成果。這通常可在第一次讀取時完成，但有時需要更深入的鑑識才能確定，我們知道差異為何。

透過在子文件層級處理輸出管道來降低延遲，以便能更快開始進行具有時效性的下游處理作業。您找到所需的資訊後，若文件中含有個人識別資訊 (PII)，您可以提早終止以避免處理檔案中不必要部分的額外作業，進而減少 CPU 使用量。

磁碟與記憶體使用量可能是一大考量，尤其是部署在端點上時。OpenText File Content Extraction 可將各部分的使用量降至最低，包括處理大型檔案時的剩餘記憶體。如果特定使用案例不需要特定功能，您可以透過部署佔用較少空間的 OpenText File Content Extraction 來進一步最佳化您的應用程式。

若效能是關鍵要素，則可選擇經過專業維護，並透過部署至數百萬名終端使用者獲得實證的解決方案。

專為 OEM 設計

獨特的應用程式對所嵌入的技術有獨特的需求。我們瞭解這一點，因此提供適合多種部署環境 (端點、伺服器/雲端和混合模型) 的多種選項，且在 Intel 和 ARM 架構上支援 Windows、Linux 和 macOS，並提供多種程式語言的原生 API。

可輕易整合的程式庫與參考程式碼，可與全新或現有的應用程式快速整合。OpenText File Content Extraction 支援檔案式和串流式 I/O，能為任何應用程式架構提供最適合的選項。

OpenText File Content Extraction 採用安全設計，透過多種技術來將風險降至最低，例如減少攻擊面以降低威脅的影響、隔離元件以快速緩解協力廠商弱點、減少程序權限，以及針對 Zip Slip 等特定格式型攻擊的對策。用於遏制威脅的安全線串、程序內或程序外模型，可改善處理任何輸入資料時的應用程式穩定性。我們瞭解您產品最大的安全威脅之一，就是您的供應鏈。

彈性的授權方式可確保我們能配合您的商業模式。

OpenText File Content Extraction 可與 OpenText Named Entity Recognition 等其他嵌入式 OpenText 技術合作無間，同時使用時可提供額外功能和效能增益。

OpenText File Content Extraction 是為了內嵌使用而存在，我們已為 OEM 市場服務多年，也瞭解您所面臨的挑戰。

主要功能

檔案格式偵測

快速準確地辨識檔案類型，以降低不當處理重要資訊，或浪費寶貴的 CPU 時間在不相關檔案上的風險。OpenText File Content Extraction 不完全依賴可偽造的檔案名稱副檔名或簡短的神奇數字，而會檢視每個檔案，將重點先放在最能明顯區分的特徵，並深入所需深度以解決不明確情況，進而加快解答速度並降低錯誤率。

OpenText File Content Extraction 除了能處理 MIME 類型，還能清晰地辨識不存在或不明確的 MIME 類型 (例如應用程式/八位元組串流)、新增加密狀態、格式分類及格式版本等詳細資料，讓您能做出準確的下游輪遞及處理決策。

支援的文件類別包含分析、動畫、CAD、資料庫、桌面發佈、封裝、可執行檔、字型、GIS、程式庫、影片、物件模組、外框、簡報圖形、點陣影像、排程、科學、聲音、原始碼 (包含語言辨識)、試算表、向量圖形與文字處理。

權限管理

辨識權限管理受到保護的 Microsoft、Seclore 及 SmartCipher 檔案。檢查 MSIP (Microsoft Purview Information Protection) 標籤 (即使是來自加密檔案) 以正確判斷風險。

解密*受 Microsoft Azure Rights Management (RMS) (屬於 Microsoft Information Protection 的一部分) 保護的檔案及相關技術，讓您的工作流程能在原始未加密的內容上透明地運作。

中繼資料存取

快速存取檔案中繼資料 (例如 XMP、XrML、IPTC、EXIF、Boldon-James 分類)，並將特定欄位格式化。OpenText File Content Extraction 可將一般欄位合併及進行標準化，讓下游能更輕鬆地使用。

字元集轉換

為通常會要求使用 UTF-8 輸入的下游程序進行準備。OpenText File Content Extraction 會自動判斷文件中使用的字元集 (即使這並非中繼資料中指定的字元集)，並將其轉換為 UTF-8 或您選擇的其他編碼。只要正確辨識並轉換，就能保留價值。

文字擷取

快速移除格式過渡支架和其他雜訊以擷取純文字內容。

深入探索多種文件格式、擷取本文和其他可見元件 (例如：頁首、頁尾、註腳、註解、標題和表格元件)，並可選擇包含隱藏的文字 (例如區段名稱、附註、追蹤修訂、明確隱藏的元素、協助工具層和可設定的預留位置文字)，以及快取、孤立、未使用和已刪除的文字。

* 需要身分證明

子檔案擷取

深入挖掘通常會內嵌更多內容的格式，從顯而易見的歸檔格式和電子郵件儲存，到更令人驚訝的容器格式 (例如 PDF 及其不同版本)，以及所有三大辦公室類型文件：文字處理、試算表和簡報圖表。

OCR

在處理管道中直接存取掃描文件、拍攝的收據，以及含有文字的點陣影像中的文字內容。

結構化輸出

您可從文件內容中擷取結構，建立格式良好且經預先定義文件類型定義 (DTD) 驗證的 XML。OpenText File Content Extraction 會將 XML 詞彙套用至文件中的資料結構，讓下游應用程式能依脈絡存取內容。

OpenText File Content Extraction 會傳回結構，例如頁首/頁尾、註腳、註解、書籤、標題、工作表名稱和結構化表格資料。

HTML 輸出

以高精確度的 HTML 預覽文件。將此技術整合至您的 Web 應用程式中，可讓您的最終使用者即使沒有適當的外掛程式或原生應用程式也可檢視文件。

透過 HTML 輸出，您可以使用能輕鬆自訂的範本或靈活健全的 API 來控制 HTML 輸出的內容、結構和格式。選擇適合網頁的動態流動文字，以妥善瞭解文件內容，或是固定寬度轉譯來模擬列印輸出。

將內容拆解成可管理的區塊，以加快載入時間並降低瀏覽器記憶體佔用空間。新增結構 (例如使用自訂標記醒目提示)，並根據文件內容 (例如字型大小或樣式) 自動產生可導覽的目錄。套用階層式樣式表 (CSS) 以改善輸出逼真度，並使外觀與風格一致，以便快速且輕鬆地閱讀。

PDF 輸出

將檔案歸檔為 PDF 格式，確保文件內容可在未來幾年凍結及存取，而不需下載特定應用程式來顯示每個檔案類型。

系統要求

	API					平台						
	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
檔案格式偵測	•	•	•	•	•	•	•	•	•	•	•	•
權限管理[4]	•	•	•	•	•	•	•	•	•	•	•	•
中繼資料存取[5]	•	•	•	•	•	•	•	•	•	•	•	•
字元集轉換	•	•	•	•	•	•	•	•	•	•	•	•
文字擷取	•	•	•	•	•	•	•	•	•	•	•	•
子檔案擷取	•	•	•	•	•	•	•	•	•	•	•	•
OCR	•	•	•	•	•		•		•	•		•
結構化輸出	•			•		•	•		•	•	•	•
HTML 輸出	•			•		•	•		•	•	•	•
PDF 輸出	•						•		•			

- 可用

進一步瞭解 https://www.opentext.com/products/idol-keyview ›

1 Python 僅在 x86_64 平台和 macOS/ARM 提供

2 .NET 僅在 Windows 平台提供

3 僅 C API 支援 Windows/ARM

4 解密 (可個別授權) 僅在 Windows/x86_64 和 Linux/x86_64 平台，以 C 與 Java API 形式提供

5 .NET 中繼資料存取不包括標準化

Copyright © 2024 Open Text • 11.24 | 260-TW0154-003

opentext™