**Notable supported file types for identification and content extraction**

For complete list of file formats identified (1,330+) and supported levels of content extraction see Open Discover Supported Formats for Version 2019.2.pdf.

|  |  |
| --- | --- |
| **File Type Classification** | **File Types** |
| Archive | 7Z, ZIP, RAR, RAR5, TAR, XAR, BZ2, Z, ARJ, CAB, GZIP, MS BINDER, HQX, MSO, AppleSingle1, AppleSingle2   * Split archive formats 7Z, ZIP, RAR are supported * 7Z, ZIP, RAR are supported for password decryption * Self-extracting 7Z, ZIP, RAR executables are also supported |
| Mail Store | PST, OST, OST 2013, MBOX, DBX |
| Media Image | WIM, ISO, HFS, HFS+, DMG, UDF, VHD, VDI, QCOW, VMDK 1-5   * Split WIM media image format is supported |
| Email | MSG, EML, EMLX, TNEF, DXL, ICS, VCF, P7M, P7S, TextMail   * S/MIME digitally-signed and encrypted emails are supported * TextMail is email saved as text (e.g., via Outlook email client). See “Text” file type classification for the various encodings that are recognized for TextMail |
| Word Processing | DOC, DOCX, DOCM, DOTX, DOTM, ODT, OTT, WPS, HWP3, HPW5, Apple iWork Pages ’05-’09, Ichitaro 5-8   * Word saved as 2003, 2007 XML is supported * Ichitaro 8 Compressed is not yet supported |
| Presentation | PPT, PPTX, PPTM, PPSX, POTX, PPSM, ODP, OTP, SHOW (Hanshow), Apple iWork Keynote ’05-’09   * PowerPoint formats 97-2003 and 2007-2016 are supported for password decryption\* * PowerPoint saved as XML is supported * Open Document (LibreOffice/OpenOffice) formats with encryption (Blowfish/AES) are supported for password decryption |
| Spreadsheets | XLS, XLSX, XLSB, XLSM, XLTX, XLTM, XLAM, CELL (Hancell), ODS, OTS, Apple iWork Numbers ’05-’09   * Excel formats 97-2003 and 2007-2016 are supported for password decryption\* * Open Document (LibreOffice/OpenOffice) formats with encryption (Blowfish/AES) are supported for password decryption |
| Raster Image | JPG, TIFF, PNG, GIF, ARW, CRW, NEF, RW2, ORF, WEBP |
| Vector Image | PDF, PDF Portfolio, PDF XFA, PDF AcroForm, XPS, VSD, VSDX, VSDM, VSSX, VDX, ODG, OTG, ODC, WMF, WMZ, EMF, EMZ   * PDF types are supported for password decryption\* * Open Document (LibreOffice/OpenOffice) formats with encryption (Blowfish/AES) are supported for password decryption |
| Multimedia | MP3, MP4, WMV, WEBM, MOV |
| Markup | HTML, XHTML, HTM, XML, MHT, RTF   * Metadata extraction supported * Embedded item extraction supported for MHT and RTF * MS Office 97-2003, 2007+ documents saved as HTML, MHT, or XML are identified as such and text and metadata are extracted (e.g., see file format Id.PowerPointMhtml and Id.PowerPointXml) |
| Notes and Research | OneNote 2010, 2013, and 2016   * OneNote 2007 is not yet supported; however, embedded item extraction and binary-to text is supported. * Password decryption for OneNote 2010-2016 is not yet supported |
| Text | Supported encodings for identification and extraction:  ASCII, UTF-7, UTF-8, UTF-16LE, UTF-16BE, UTF-32LE, UTF-32BE, 1250, 1251, 1253, 1254, 1255, 1256, 8859-1, 8859-2, 8859-5, 8859-6, 8859-7, 8859-8, 8859-9, KOI8-R, SHIFT-JIS, EUC-JP, ISO-2022-JP, EUC-KR, ISO-2022-KR, ISO-2022-CN, Big5, GB18030, IBM 424, IBM 420, IBM 866, EBCDIC 500   * All text extracted is converted to a .NET Unicode string |
| Project Management | MPP: Microsoft Project 97-2003, 2007-2016 supports metadata, embedded item extraction, and limited text extraction |
| Database | Domino XML document database (.dxl;.xml) |

\*DRM or custom encryption is not supported. Excel and PowerPoint 97-2003 formats that are protected with default password are automatically decrypted, no password is required.

In addition to the above notable supported file formats, Open Discover SDK offers a binary-to-text content extractor for formats that are not supported. The binary-to-text content extractor “scrapes” out text in UTF-8, UTF-16, and code page 1252 encodings. UTF-8 “scraping” supports all language code ranges while UTF-16 and code page 1252 only support “Latin” encodings. In many cases useful text for indexing can be extracted via binary-to-text extraction.