



V 5 . 5

# 利用ガイド

1.0 版

2020 年 08 月 20 日



改訂履歴

2020 年 08 月 20 日

V5.5 用の初版

## 目次

1. はじめに.....	6
ユーザサポートについて .....	7
ユーザサービスについて .....	8
2. 機能概要.....	9
2.1 特徴.....	9
2.2 対応プラットフォーム.....	9
2.3 機能紹介 .....	10
2.4 共通の制限事項.....	16
2.5 対応文字言語と文字コード .....	16
3. 各公開 API とインターフェース部の処理 .....	18
3.1 機能概要 .....	18
3.2 型定義 .....	18
3.3 エラーメッセージ .....	19
3.4 公開 API 詳細 .....	22
3.4.1 ファイル識別関数 .....	23
3.4.2 テキスト抽出関数 .....	36
3.4.3 テキスト抽出関数（ストリーム出力） .....	46
3.4.4 プロパティ抽出仕様 .....	47
3.4.5 頁抽出関数.....	57
3.4.6 頁抽出関数（ストリーム出力） .....	59
3.4.7 パスワード付きファイルのテキスト抽出関数.....	60
3.4.8 パスワード付きファイルの頁抽出関数.....	60
3.4.9 パスワード付きファイルのプロパティ抽出関数.....	61
4. テキスト抽出仕様.....	62
4.1 共通仕様 .....	62
4.1.1 制御コード.....	62
4.1.2 定義外文字.....	62
4.1.3 ユーザ外字.....	62
4.1.4 OLE オブジェクト抽出 .....	62
4.1.5 圧縮ファイルからの抽出 .....	65
4.1.6 ストリームへの抽出.....	65
4.2 ワープロ文書 .....	65
4.2.1 全角文字罫線（一太郎、OASYS、OASYS オンラインのみ有効） .....	65

4.2.2 その他の罫線 .....	65
4.2.3 表 .....	65
4.2.4 RTF .....	66
4.2.5 一太郎 8 から一太郎 13、一太郎 2004 から一太郎 2019 .....	67
4.3 プレゼンテーションファイル抽出仕様 .....	67
4.3.1 テキスト抽出処理概要 .....	67
4.3.2 抽出データ中のタグ出力 .....	67
4.4 表計算 .....	68
4.5 PDF .....	70
4.6 CAD .....	71
4.7 HTML .....	71
4.8 XML .....	71
4.9 OFFICE 2001 FOR MACINTOSH, EXCEL98 FOR MACINTOSH .....	72
4.10 QUARKXPRESS .....	72
4.11 DOCUWORKS .....	72
4.12 VISIO .....	73
4.13 OUTLOOK/OUTLOOK EXPRESS .....	73
4.14 OFFICE 2007/OFFICE 2010/OFFICE 2013/OFFICE 2016/OFFICE2019 .....	73
4.15 OASYS .....	74
4.16 OPENOFFICE 1.0 .....	74
4.17 OPENOFFICE.ORG 3.1/3.2/3.3, LIBRE OFFICE 3.3/3.4 .....	74
5. 付録 .....	75
5.1 開発環境 .....	75
5.2 TEXT_OEM.H のプラットフォーム定義マクロについて .....	75
5.3 著作権情報 .....	78
5.4 サンプルアプリケーションの使用法 .....	82
1. 入力アプリケーションファイルの指定 .....	82
2. 抽出先ファイル名の指定 .....	83
3. 抽出先ファイルの格納場所の指定 .....	83
4. オプション .....	83
5.5 To_COM_VCS の使用法 .....	88
5.6 LIVIEW の使用法 .....	91
5.7 パスワード付き PDF 文書のテキスト抽出 .....	93
5.8 セキュリティ設定した PDF のテキスト抽出制御仕様 .....	95
5.9 パスワード付き MICROSOFT OFFICE、一太郎のテキスト抽出 .....	101



# 1. はじめに

---

この度は「TextPorterV5.5 Server 版」をお求め頂き、ありがとうございます。  
本ライブラリは、PDF、Microsoft Office、一太郎などのファイルからテキスト、プロパティ情報の抽出機能を持つライブラリです。

本製品は以下のライセンスを用意しております。  
別途契約書に記載されている使用許諾内容を守って頂くことが前提になります。

## 通常ライセンス

官公庁・企業・団体等のサーバアプリケーションまたは、システムへ組み込みご利用いただくためのライセンスです。

## デベロッパライセンス

開発用のシステムのみで使用する事ができるライセンスです。実運用のシステムでは使うことができません。

## OEM ライセンス

アプリケーションに組み込んで、アプリケーションと共に複製・頒布する権利を提供するライセンスです。

本利用ガイドに記載の商品名は各社の商標または登録商標です。

本利用ガイドに記載の内容一部または全部を無断で複写・転載することを禁じます。

本利用ガイドに記載の内容及び製品の仕様などは予告無く変更されることがあります。

## ユーザサポートについて

### ■ ユーザサポート提供内容

ユーザサポート専用の窓口で次のようなユーザサポートをご提供します。

- ① ご購入後 1 年間限り、製品機能、利用方法等に関するご質問にお答えします。
- ② 契約日から 1 年以内に重大な障害が発見された場合は、無償で修正版をご提供致します。
- ③ ご購入後 1 年以内にご購入製品のバージョンアップが行われた場合は、無償でご提供します。
- ④ 契約日から 1 年を経過した場合、別途、保守契約を結んで頂いて、前年度同様のサポートを提供いたします。

### ■ ユーザサポート窓口

本製品に関するお問い合わせは下記あてにメールでお願い致します。

お問い合わせの際は、必ず製品のシリアル番号をご提示ください。

アンテナハウス株式会社 G2グループ

E-mail:dmc-support@antenna.co.jp

〒399-4511 長野県上伊那郡南箕輪村字堀北 8077-1

## ユーザサービスについて

### ■ ユーザサービス内容

万一、CD-ROM を破損された場合は、2,500 円（消費税別）で新しい CD-ROM とお取り替えます。製品のシリアル番号を添えてお申し込みください。

郵便振替にて下記宛にお申し込みください。

振替番号：02長野ー00590ー2ー5175 アンテナハウス株式会社
---------------------------------------

.....  
☞ お申し込み内容・お名前・ご住所・シリアル番号を必ずご記入ください。  
.....



## 2. 機能概要

---

### 2.1 特徴

本ライブラリは、次のような設計思想の元に開発されたライブラリです。

- テキスト抽出に最適化されたコンパクトかつ高速なモジュール
- 新しいアプリケーションファイル形式に迅速に対応できるフレキシブルな構造
- 日本の国内外を問わず幅広いアプリケーションから利用可能な多言語対応モジュール
- サーバはもとより、デスクトップ PC、携帯情報端末等あらゆる環境で動作可能なスレッドセーフかつポータブルなモジュール

### 2.2 対応プラットフォーム

次の環境で動作するライブラリを用意しています。アプリケーションから呼び出すためのライブラリのインターフェイスは共通です。

- Microsoft Windows

Windows 8.1(32bit/64bit)/Windows 10(32bit/64bit)

Windows Server 2012(64bit)/Windows Server 2012 R2(64bit)

Windows Server 2016(64bit)

Windows Server2019(64bit)

\* Microsoft Visual C++ 2019 でビルドしています。

動作には、Visual C++ 2015 から 2019 に共通のランタイムが必要です。システムにインストールされていない場合は、製品パッケージの `redist` フォルダにある「Microsoft Visual C++ 2019 再頒布可能パッケージ」をインストールしてください。

32bit 版は `VC_redist.x86.exe`, 64bit 版は `VC_redist.x64.exe` です。

- Linux (32bit/64bit)

\* GCC 8.3.1 でビルドしています。

動作には、`libc-2.28.so`, `libstdc++.so.6` 以上が必要です。

動作保証については、対応プラットフォーム(OS, JavaVM など)に起因する問題は、弊社では保証できません。また、プラットフォームに起因する問題に対する解決先・回避策の提供は、通常サポートには含まれません。

TextPorter の販売中、あるいは有償保守契約の期間中であっても、プラットフォーム製造元のサポート期間が終了した場合、動作保証はできません。

プラットフォーム製造元のサポート期間が終了したあとも、TextPorter の動作保証をお求めの場合は、弊社までご相談ください。

## 2.3 機能紹介

各機能を紹介します。

なお、当該アプリケーション以外で作られた互換ファイルについては、動作保証ができません。

### ファイル識別機能

ファイルを作成したアプリケーション名とそのバージョンを識別します。ファイルの拡張子ではなく、ファイル内部の情報に基づいて識別しますので、正確な判別が可能となります。識別可能なファイルは以下の通りです。

表 1 識別対象ファイル形式一覧

アプリケーション	ファイル形式	拡張子
PDF 1.0 から 1.7		PDF
Microsoft RTF		RTF
Microsoft Word 2007 Microsoft Word 2010 Microsoft Word 2013 Microsoft Word 2016/2019	ネイティブ形式 マクロ対応形式 テンプレート形式 マクロ対応テンプレート形式	DOCX DOCM DOTX DOTM
Microsoft Word 2000/XP/2003 Microsoft Word Ver.6/7(95)/97/98 Microsoft Word 2001 for Macintosh	ネイティブ形式	DOC (変更可)
Microsoft Excel 2007 Microsoft Excel 2010 Microsoft Excel 2013 Microsoft Excel 2016/2019	ブック形式 マクロ対応ブック形式 テンプレート形式 マクロ対応テンプレート形式	XLSX XLSM XLTX XLTM
Microsoft Excel 2000/XP/2003 Microsoft Excel Ver.4/5/7(95)/97 Microsoft Excel 98 for Macintosh Microsoft Excel 2001 for Macintosh	ブック形式	XLS
Microsoft PowerPoint 2007 Microsoft PowerPoint 2010 Microsoft PowerPoint 2013 Microsoft PowerPoint 2016/2019	プレゼンテーション形式 マクロ対応プレゼンテーション形式 テンプレート形式 マクロ対応テンプレート形式 スライドショー形式 マクロ対応スライドショー形式	PPTX PPTM POTX POTM PPSX PPSM
Microsoft PowerPoint 2000/XP/2003 Microsoft PowerPoint 95/97 Microsoft PowerPoint 2001 for Macintosh		PPT 及 び PPS
Microsoft Visio2000/2002/2003/2007/2010 Visio V3/V4/V5		VSD

Microsoft Visio2013		VSDX VSDM VSSX VSSM VSTX VSTM
Outlook 2000/2003/2007/2010 Outlook Express 6		eml 及 び msg
－太郎 2004 から 2019 －太郎 Lite/9 から 13 －太郎 8/R1 for Windows95/NT －太郎 8 Office Edition R.1 －太郎 8 Office Edition R.2		JTD 及 び JTT
－太郎 7 /R1 for Windows －太郎 7 /R2 for Windows		JFW 及 び JVW
－太郎 Ver.6/R1 for Windows －太郎 Ver.6.3/R1 for Windows －太郎 Dash2 for Windows		JBW 及 び JUW
－太郎 Ver.5 －太郎 Ver.5/R.2 for DOS/V －太郎 Ver.5/R.1 for Windows －太郎 Ver.5/R.2 for Windows		JAW
－太郎 dash/Ver.4/Ver.4.3		JSW
OASYSV10 OASYS2002 OASYS V5/V6/V7V8	分離型ファイル 結合型ファイル 複合型ファイル	FMT、DOC OA2 OA3
OASYS for Windows Ver.4.0 OASYS for Windows Ver.4.1	分離型ファイル 結合型ファイル	FMT、DOC OA2
OASYS for Windows Ver.3.0 OASYS for Windows Ver.3.0a	分離型ファイル 結合型ファイル	FMT、DOC OA2
OASYS1 形式 OASYS V5/V6 のオンライン形式		OAS
Lotus 1-2-3 R5J		WK4
Lotus 1-2-3 97/98/2000 Lotus 1-2-3 MillenniumEdition		123
HTML		HTM or HTML
TEXT		任意指定
AutoCAD GX-III DXF AutoCAD GX-5/R12 DXF AutoCAD R13/R14 DXF AutoCAD 2000 から 2012 DXF		DXF
AutoCAD GX-III DWG AutoCAD GX-5/R12 DWG AutoCAD R13/R14 DWG AutoCAD 2000 から 2012 DWG		DWG
IGES		IGS
Adobe PageMaker 6.0/6.5/7.0		PMD PM6 P65 及 び T65
ClarisWorks 4.0 (WP、SS、DB)		CWJ
AppleWorks 6(WP、SS、DB、DR、PR、PT)		なし

XML Word XML (Word2003 保存した XML) Excel XML (Excel2003 保存した XML)		XML
WordPerfect 7/8/9		WPD
Mac Write II		なし
Microsoft Works 2000 (WP) Microsoft Works 2000 (SS) Microsoft Works 2000 (DB)		WPS WKS WDB
Corel Presentations 9		SHW
QuarkXpress 3.3 / 4		QXD
DocuWorks v3/v4/v5/v6/v7/v8		XDW
OpenOffice1.0 Calc Draw Global Impressi Math Writer		SXC SXD SXG SXI SXM SXW
OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Calc Presentation Writer		ODS ODP ODT
WPS Office 97/2000/2002/2003	金山文字形式	WPS
圧縮ファイル LZH ZIP RAR		LZH ZIP RAR

### 【制限事項】

#### ① 他のバージョンに含まれるフォーマット

以下のフォーマットは個別に識別できないため、他のバージョンのカテゴリに含まれます。

- Microsoft ExcelXP、Excel2000 と Excel97 は区別不可のため、同じカテゴリとして識別されます。
- Microsoft Excel Ver.5 と Microsoft Excel Ver.7(95)は区別不可のため、同じカテゴリとして識別されます。
- PowerpointXP、Powerpoint2000 と Powerpoint97 は区別不可のため、同じカテゴリとして識別されます。
- 一太郎 Lite、一太郎 8 から一太郎 13、一太郎 2004 から一太郎 2019 は同じカテゴリとして識別されます。
- 一太郎 Dash2 は一太郎 Ver.6 と同じカテゴリとして識別されます。
- OASYS1-2-3 は Lotus1-2-3 97/98 と同じカテゴリに識別されます。
- WordPerfect 7、WordPerfect 8、WordPerfect 9 は同じカテゴリとして識別されます。
- AutoCAD R13 と LT95 は区別不可のため、同じカテゴリとして識別されます。
- AutoCAD R14 と LT97 は区別不可のため、同じカテゴリとして識別されます。
- AutoCAD 2000 と AutoCAD 2002/2004/2005/2006 は、同じカテゴリとして識別されます。
- AutoCAD 2007/2008/2009 は、同じカテゴリとして識別されます。

- AutoCAD 2010/2011/2012 は、同じカテゴリとして識別されます。
- DocuWorks v4 と DocuWorks v5/v6 は区別不可のため、同じカテゴリとして識別されます。  
DocuWorks v8 は、DocuWorks v7 として識別されます。
- PageMaker v6.5 と PageMaker v7.0 は区別不可のため、同じカテゴリとして識別されます。
- Visio 2007, Visio 2010 は Visio 2003 と同じカテゴリとして識別されます。

## ② OASYS のバージョンの識別

OASYS 文書は、バージョンを特定できない場合があります。その場合、adSubFormat は設定されません。

## ③ OASYS の分離形式の識別方法

OASYS 分離形式は、\*.FMT 文書を参照して識別するため、同じフォルダに\*.FMT ファイルがあることが条件となります。

## ④ Lotus1-2-3 から Excel 形式で保存したファイル

Lotus1-2-3 2000 または MillenniumEdition から Excel97 形式で保存したファイルは、作成環境の情報が取得できないため、Excel98 と識別されます。

## ⑤ 暗号化された Microsoft Office ファイル

暗号化された Microsoft Office ファイルは、パスワードを指定しない API では、正確に判別できないため、ファイル種別は、「Protected Microsoft Office Document」となり、error 3001 を返します。

## ⑥ 暗号化された Microsoft Office ファイルの OLE と添付

暗号化された Microsoft Office ファイルが、その中に OLE で埋め込まれたものを含む場合、OLE で埋め込まれたものは、抽出できません。

暗号化された Microsoft Office ファイルが、他のアプリケーションに添付された場合は、抽出できません。

## ⑦ OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4 の暗号化されたファイル

OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4 の暗号化されたファイルは、error3000 を返します。

## ⑧ PDF

Acrobat X で作成した 256-bit AES 形式の暗号には、対応していません。error5000 を返します。

## テキスト抽出機能

文書を作成したアプリケーションが無くても、指定したファイル、またはファイルに埋め込まれた OLE オブジェクトからテキスト文字列を取り出せます。

対応アプリケーションは以下の通りです。

表 2 抽出対象ファイル形式一覧

※対応アプリケーション・バージョン・言語・OLE 抽出は「○」を示す。

アプリケーション	Japanese	English	GB18030	GBK	BIG5	KS_C	OLE In	OLE Out
<u>Microsoft</u>								
Word Ver6/Ver7(95)/97/98/2000/2001 for Mac/XP/2003	○	○	○	○	○	○	○	○
Word 2007/2010/2013/2016/2019	○	○	○	○	○	○	○	○
Excel Ver4/Ver5/Ver7(95)	○	○	○	○	○	○	○	○
Excel 97/98 for Mac/2000/2001 for Mac/XP/2003	○	○	○	○	○	○	○	○
Excel 2007/2010/2013/2016/2019	○	○	○	○	○	○	○	○
PowerPoint 95/97/2000/2001 for Mac/XP/2003	○	○	○	○	○	○	○	○
PowerPoint 2007/2010/2013/2016/2019	○	○	○	○	○	○	○	○
RTF	○	○	○	○	○	○	○	
Works 2000 WP (文書)	○						○	
Works 2000 SS (表)	○							○
Works 2000 DB (データベース)	○						○	
Visio 2000/2002/2003/2007/2010	○						○	○
Outlook2000/2003/2007/2010(*.msg形式)	○	○	○	○	○	○	○	
Outlook Express 6(*.eml形式)	○	○	○	○	○	○	○	
<u>Visio</u>								
VisioV4/V5	○						○	○
Visio 2013	○	○	○	○	○	○	○	○
<u>JUSTSYSTEM</u>								
一太郎 Ver5/Ver6/Ver6.3/Dash2	○							
一太郎7							○	
一太郎Lite、8から13、2004から2019	○						○	○
<u>Adobe</u>								
Adobe PageMaker 6.0/6.5/7.0	○	○	○	○	○	○	○	
<u>PDF</u>								
PDF 1.2から1.7	○	○	○	○	○	○		○
<u>富士通</u>								
OASYS V3/V4/V5/V6/V7/V8/2002/V10 分離形式	○							
OASYS V3/V4/V5/V6/V7/V8/2002/V10 結合型(OA2)形式	○							
OASYS V5/V6/V7/V8/2002/V10 複合型(OA3)形式	○							
OASYS Online形式	○							
<u>Lotus</u>								
Lotus1-2-3 R5	○	○	○	○		○		
Lotus1-2-3 97/98/OASYS 1-2-3 V6/V7/V8/2000	○	○	○	○		○		
Lotus1-2-3 MillenniumEdition9.5		○						

アプリケーション	Japanese	English	GB18030	GBK	BIG5	KS_C	OLE In	OLE Out
<u>Claris</u>								
Mac Write II	○							
ClarisWorks 4.0	○							
<u>Apple</u>								
AppleWorks 6	○	○						
<u>Corel</u>								
WordPerfect Office 2000 (WordPerfect 8/9 のみ)	○		○	○				
Corel Presentations 9(Slide show 7/8/9)	○	○						
<u>Quark</u>								
QuarkXPress 3.3/4	○							
<u>AutoDesk</u>								
AutoCAD R13(LT95)/R14(LT97) DXF Ascii形式	○	○	○	○	○	○	○	
AutoCAD R13(LT95)/R14(LT97) Binary形式	○	○	○	○	○	○	○	
AutoCAD 2000から2012 DXF	○	○	○	○	○	○	○	
AutoCAD R13(LT95)/R14(LT97) DWG	○	○	○	○	○	○	○	○
AutoCAD 2000から2012 DWG	○	○	○	○	○	○	○	○
IGES	○	○						
<u>Fuji XEROX</u>								
DocuWorks v4/v5/v6/v7/v8	○	○					○	
<u>OpenOffice.org</u>								
OpenOffice 1.0								
OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4	○	○	○	○	○	○		○
<u>金山ソフト</u>								
WPS Office 97/2000/2002/2003 (金山文字のみ)			○	○	○		○	○
<u>その他</u>								
HTML	○	○	○	○	○	○		
XML/ Word XML/ Excel XML	○	○	○	○	○	○		

※OLE In : 他のアプリケーション文書を OLE として文書に挿入したものを抽出できます。

※OLE Out : 文書を OLE として、他のアプリケーション文書に挿入したものを抽出することができます。

OLE に関しては、○が付いていても、制限がある場合があります。詳しくは、「4. テキスト抽出仕様」を参照してください。

※LZH, ZIP, RAR の3種類の圧縮ファイル(アーカイブファイル)については、表にあるテキスト抽出対象ファイルが含まれていた場合に、抽出が可能。

### プロパティ抽出機能

指定したファイルから「プロパティ」情報を抽出し、プロパティ構造体に設定します。

### 頁抽出機能

アプリケーションファイルから指定された頁のテキストを抽出します。

### パスワード付きファイルのテキスト抽出機能

ユーザパスワードが設定された PDF ファイルや、パスワードで保護された(暗号化され

た)Microsoft Office ファイルからテキスト文字列、プロパティ情報、頁のテキストを抽出します。

#### ストリーム抽出機能

指定したファイル、または、ファイルに埋め込まれた OLE オブジェクトからテキスト文字列をストリームへ取り出します。

ストリーム抽出機能がない言語インターフェースもあります。

## 2.4 共通の制限事項

抽出元アプリケーションに対応するテキスト抽出エンジンと、その使用ライセンスがないものは、テキスト抽出できません。

PDF ファイルと Microsoft Office ファイルを除いて、パスワードで保護した(暗号化された)ファイルからのテキスト抽出はできません。パスワード保護されたファイルはあらかじめオリジナルのアプリケーションでパスワード設定を解除する必要があります。

Microsoft の IRM (InformationRightmanagement) 機能を使って、ドキュメントへのアクセスの制限 (閲覧、変更 Etc.) を設定した Microsoft office ファイルからのテキスト抽出はできません。

当該アプリケーション以外で作られた互換ファイルについては、動作保証ができません。

ファイルに存在するテキストを全て抽出できるとは限りません。また、アプリケーションで表示されるテキストをすべて抽出できるとは限りません。TextPorter の仕様上、抽出対象になってないテキストが存在する可能性があります。

本製品の仕様は、改良のため予告なく変更することがあります。製品の改良に伴い、テキスト抽出結果が以前と異なる結果になる場合があります。

## 2.5 対応文字言語と文字コード

本ライブラリでは、抽出元アプリケーションにて様々な方式で符号化された文字集合データからテキストを抽出します。抽出対象文字列は、Unicode 文字集合体、日本語、英語(ラテンアルファベット文字集合 ISO\_8859-1 のみ)、中国語 (簡体字 GB18030、GBK、繁体字 Big5)、韓国語 (KS\_C\_5601\_1987) の文字列です。上記以外の言語のフォントを使用している個所の抽出結果は保証できません。

抽出先の文字集合については、アプリケーションから下記の符号化方式を任意に指定可能です。

表 3 抽出テキストの文字符号化方法

文字符号化方式	基本文字集合	ユーザ外字領域
①Shift_JIS	JIS-X0201,X0208	なし
②WINDOWS31J	Microsoft Win3.1J 文字セット	1880 文字



③EUC-JP	JIS-X0201,X0208,X0212	なし
④EUC-JP-FIX		
⑤ISO-2022-JP	JIS-X0201Latin,ISO646IRV,X0208,C6226	なし
⑥ISO-10646-UCS-2	16 ビット固定/文字	
⑦ISO-10646-UCS-4	32 ビット固定/文字	
⑧UTF-8		
⑨UTF-16	UCS2+サロゲートペア	
⑩ISO_8859-1		
⑪GB18030		
⑫GBK		
⑬Big5		
⑭KS_C_5601_1987	KS X 1001+8822 additional hangul	
⑮Shift_JIS-2004		
⑰ISO-2022-JP-2004		
⑱EUC-JIS-2004		

## 3. 各公開 API とインターフェース部の処理

---

### 3.1 機能概要

インターフェース機能部は、テキスト抽出ライブラリを使用するアプリケーションとライブラリ間のインターフェースを提供します。

主な機能は以下の 9 つで、これらを公開 API として実装します。

なお、V4 の API（末尾が\_V4）は、次期バージョンでは廃止しますので、すみやかに V5 の API に移行するようにお願いいたします。

- ファイル識別機能
- テキスト抽出機能
- テキスト抽出機能（ストリーム出力）
- プロパティ抽出機能
- 頁抽出機能（一部エンジン実装）
- 頁抽出機能（ストリーム出力。一部エンジン実装）
- 暗号化された PDF や Microsoft Office ファイルのテキスト抽出機能
- 暗号化された PDF や Microsoft Office ファイルのプロパティ抽出機能
- 暗号化された PDF や Microsoft Office ファイルの頁抽出機能

### 3.2 型定義

本ライブラリでは、プラットフォームの違いを吸収するために、

INT, BOOL, Byte, Word, DWord, LWord, int8, uint8, int16, uint16, int32, uint32, int64, uint64

など、さまざまな型定義を行っています。

型の詳細については、製品パッケージの Include ディレクトリにある text\_oem.h を参照してください。

### 3.3 エラーメッセージ

#define SystemError	11	システムエラー
#define MemoryNotEnough	12	メモリが不足しています
#define UserAbort	13	ユーザによる中断
/* Physical file I/O errors */		
#define FileNotFound	21	ファイルが見つかりません
#define FileCantOpen	22	ファイルがオープンできません
#define FileCantCreate	23	ファイルが作成できません
#define FileCantWrite	24	ファイルが書き込めません
#define FileCantRead	25	ファイルが読み込めません
#define FileCantDelete	26	ファイルが削除できません
#define FilePathTooLong	27	ファイル名、パス名が長すぎます
/* Interface function errors */		
#define DllLoadFailed	1001	DLL がロードできません
#define DllFuncCantFound	1002	DLL に該当機能がありません
#define CantInitialize	1003	初期化用定義データパス取得エラー
#define LicenseFileNotFond	1004	ライセンス管理ファイルが見つかりません
/* File is not supported */		
#define CantDetectFile	2001	ファイルが識別できません
#define NotSupported	2002	ファイルをサポートしていません
#define TimeOver	2003	使用期限が切れています
#define LanguageNotSupported	2004	指定した抽出先文字集合はサポートしていません
/* Can't extract text */		
#define OtherProblems	3000	その他
#define ProtectedByPassword	3001	パスワード付きファイル
#define InvalidFile	3002	ファイルの内容が異常です
#define NoTextStringFound	3003	抽出可能な文字がありません
#define StoppedByOle	3004	OLE が見つかったため、プログラムを中止します

#define FileLengthOver	3005	テキスト取りだしサイズを指定した場合、サイズ指定よりもテキストデータが多いときにこのメッセージを表示（テキスト抽出は、指定サイズで行います。）
#define MaxLoopCounts	3006	無限ループ自動検知（オプション DMC_GETTEXT_OPT_LOOP の指定）を行う際、ループ数>MAX_LOOP（0x40000）の時にこのエラーコードを返します。
#define CompressDateUncompress	3007	DocuWorks：ファイル内に圧縮されたデータがあります。 PDF：ファイルに LZW 圧縮されたテキストが有り解凍できません
/* OLE */		
#define OleCantDetectFile	4000	OLE ファイルが識別できません
#define OleNotSupported	4001	OLE ファイルをサポートしていません
#define OleProtectedByPassWord	4002	パスワード付き OLE ファイル
#define OleInvalidFile	4003	OLE ファイルの内容が異常です
#define OleNoTextStringFound	4004	OLE ファイルに抽出可能な文字がありません
#define OleOutOfLimit	4005	OLE 抽出の階層制限（3 階層まで抽出可能）
#define InsertFileOutOfLimit	4006	外部ファイル抽出の階層制限（3 階層まで抽出可能）
#define InsertFileNotSupported	4007	添付の外部ファイルをサポートしていません
#define InsertFileNoTextStringFound	4008	添付の外部ファイルに抽出可能な文字がありません
/* PWD */		
#define ProtectedByPassWordPDF	5000	ユーザパスワード付き PDF ファイル
#define WrongPassWord	5001	入力されたパスワードが違います
#define PWDfileNotSupported	5002	パスワード付きファイルをサポートしていません。
#define ProtectedByOwnerPassWordPDF	5003	①PDF Owner Password によるセキュリティ設定（パーミッション設定）がされたファイルです。 ②Word 2003、Excel 2003、PowerPoint 2003 「アクセス許可」設定でアクセスを制限して

あるファイルです。

```
/* CompressedFile */  
#define CompressedFileCantDetect      6000 圧縮ファイル中のファイルを識別できません  
#define CompressedFileNotSupported    6001 圧縮ファイル中のファイルをサポートしていません。  
  
#define CompressedFileNoTextStringFound 6002 圧縮ファイル中のファイルから抽出可能な文字列がありません  
  
#define CompressedFileInvalidfile      6003 圧縮ファイル中のファイルが異常  
#define CompressedFileByPassWord       6004 圧縮ファイル中のファイルがパスワード付き文書  
  
#define CompressedFileOutOfLimit       6005 圧縮ファイルの階層制限（3階層まで抽出可能）  
  
#define StoppedByCompressedFile        6006 DMC_GETTEXT_OPT1_COMPRESS1 の設定時、圧縮ファイルが見つかりました。
```

### 3.4 公開 API 詳細

ここでは、V5 の API についてのみ、説明します。

ヘッダファイル `text_oem.h` には、V4 の API(末尾が `_V4`)が、互換性のために残してありますが、あくまで V5 の API に移行するための経過措置ですので、すみやかに V5 の API に移行することを強く推奨します。

ヘッダファイル `text_oem.h` には、Windows 版のみで使える、末尾が `_V5W` となる API があります。これは、Appfile などファイル名を想定しているパラメータの文字集合が UTF-16 の API です。機能は、末尾が `_V5` の通常の API と同じです。末尾が `_V5W` の API を使うことで、Windows で使える国際化されたファイル名がすべて使えるようになります。

### 3.4.1 ファイル識別関数

指定されたファイルを識別します。

**INT DMC\_GetFileInfo\_V5(Byte\* Appfile, DMC\_FILEINFO\* FileInfo, DMC\_TEXTINFO\_V5\* TextInfo)**

#### 引数

- Appfile : 識別対象アプリケーションファイルパス名
- FileInfo : ファイル情報構造体アドレス
- TextInfo : テキスト情報構造体アドレス（詳細は「3.4.2 テキスト抽出関数」を参照）

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

#### 説明

- ファイル情報構造体 (DMC\_FILEINFO)

```
typedef struct{
    char        DocFormat[FORMATNAME];
    char        DocSubFormat[SUBFORMAT];
    char        DocCountry[COUNTRYNAME];
    int         ProtectCode;
    int         FileType;
} DMC_FILEINFO *LPDMC_FILEINFO;
```

DocFormat	: ファイルフォーマット文字列 (アプリケーション名&バージョン)
DocSubFormat	: ファイルフォーマット補足情報、ファイル形式の 補足
DocCountry	: ドキュメント Country レコード情報
ProtectCode	: ファイルプロテクトコード
FileType	: ドキュメントタイプ

#### ■ テキスト情報構造体 (DMC\_TEXTINFO\_V5)

ファイルの識別では、TextInfo には、通常は、NULL を指定しておけば十分です。

DocCountry が識別できないときは、TextInfo->DefLangName に指定した言語名を使用します。

デフォルトの言語名を指定したいときは、TextInfo->DefLangName を指定して、TextInfo のアドレスを渡してください。

DMC\_TEXTINFO\_V5 の詳細については、「4.3.2 テキスト抽出関数」を参照してください。



**対象ファイル形式のファイル識別結果**

プロテクトコードは各ファイルにより、DMC\_TRUE または DMC\_FALSE がセットされます。

ドキュメントタイプは以下のファイル形式においては

adWORDPROCESSOR	= 1	/* Word Processor Document	*/
adSPREADSHEET	= 2	/* Spread Sheet Document	*/
adPRESENTATION	= 6	/* Presentation Document	*/
adJWORDPRO	= 22	/* Japanese Word processor	*/

のいずれかが設定されます。

**DMC\_FILEINFO に設定される文字列の詳細****OASYS オンライン**

DocFormat : “OASYS Online”  
 DocSubFormat : なし  
 DocCountry : “Japanese”

**OASYS 分離形式**

DocFormat : “OASYS Separated”  
 DocSubFormat : “V3” “V4” “V5” “V6” “V7” “V8” “2002” “V10”  
 DocCountry : “Japanese”

**OASYS 結合形式**

DocFormat : “OASYS Combined”  
 DocSubFormat : “V3” “V4” “V5” “V6” “V7” “V8” “2002” “V10”  
 DocCountry : “Japanese”

**OASYS 複合形式**

DocFormat : “OASYS Compound”  
 DocSubFormat : “V5” “V6” “V7” “V8” “2002” “V10”  
 DocCountry : “Japanese”

**Word**

DocFormat : “Word Ver.6” “Word Ver.7(95)” “Word 97” “Word 98” “Word 2000”  
 “Word XP” “Mac-Word 2001” “Word 2003”

DocSubFormat : なし

DocCountry : “Japanese” “English” “Simplified Chinese”  
 “Traditional Chinese” “Korean”

**Word 2007**

DocFormat : “Word 2007” “Word 2007 with Macro” “Word 2007 Template”  
 ” Word 2007 Template with Macro” “Word 2007 Encrypted”

DocSubFormat : なし

DocCountry : なし

**Word 2010**

DocFormat : “Word 2010” “Word 2010 with Macro” “Word 2010 Template”  
 ” Word 2010 Template with Macro” “Word 2010 Encrypted”

DocSubFormat : なし

DocCountry : なし

**Word 2013**

DocFormat : “Word 2013” “Word 2013 with Macro” “Word 2013 Template”  
 ” Word 2013 Template with Macro” “Word 2013 Encrypted”

DocSubFormat : なし

DocCountry : なし

**Word 2016/2019**

DocFormat : “Word 2016” “Word 2016 with Macro” “Word 2016 Template”  
 ” Word 2016 Template with Macro” “Word 2016 Encrypted”

DocSubFormat : なし

DocCountry : なし

**Excel**

DocFormat : “Excel Ver.4” “Excel Ver. 5/7(95)” “Excel 98” “Excel 97/2000/XP”  
 “Mac-Excel 2001” “Excel 2003”

DocSubFormat : なし

DocCountry : “Japanese” “English” “Simplified Chinese”  
 “Traditional Chinese” “Korean”

### **Excel 2007**

DocFormat :” Excel 2007” “Excel 2007 with Macro” “Excel 2007 Template”  
 ” Excel 2007 Template with Macro”

DocSubFormat : なし

DocCountry : なし

### **Excel 2010**

DocFormat :” Excel 2010” “Excel 2010 with Macro” “Excel 2010 Template”  
 ” Excel 2010 Template with Macro”

DocSubFormat : なし

DocCountry : なし

### **Excel 2013**

DocFormat :” Excel 2013” “Excel 2013 with Macro” “Excel 2013 Template”  
 ” Excel 2013 Template with Macro”

DocSubFormat : なし

DocCountry : なし

### **Excel 2016/2019**

DocFormat :” Excel 2016” “Excel 2016 with Macro” “Excel 2016 Template”  
 ” Excel 2016 Template with Macro”

DocSubFormat : なし

DocCountry : なし

### **一太郎**

DocFormat : “Ichitaro V4” “Ichitaro V5” “Ichitaro V6” “Ichitaro V7”  
 “Ichitaro Document”

DocSubFormat : なし

DocCountry : “Japanese”

### **PDF**

DocFormat : “PDF1.0” “PDF1.1” “PDF1.2” “PDF1.3” “PDF1.4” “PDF1.5” “PDF1.6”  
 ”PDF1.7”

DocSubFormat :なし

DocCountry :なし

### **PowerPoint**

DocFormat : “PowerPoint 95” “PowerPoint 97/2000/XP”  
 “Mac-PowerPoint 2001” "PowerPoint 2003"

DocSubFormat :なし

DocCountry :なし

### **PowerPoint 2007**

DocFormat : “PowerPoint 2007” “PowerPoint 2007 with Macro”  
 : “PowerPoint 2007 Slide Show”  
 : “PowerPoint 2007 Slide Show with Macro”  
 “PowerPoint 2007 Template”  
 ”PowerPoint 2007 Template with Macro"

DocSubFormat :なし

DocCountry :なし

### **PowerPoint 2010**

DocFormat : “PowerPoint 2010” “PowerPoint 2010 with Macro”  
 : “PowerPoint 2010 Slide Show”  
 : “PowerPoint 2010 Slide Show with Macro”  
 “PowerPoint 2010 Template”  
 ”PowerPoint 2010 Template with Macro"

DocSubFormat :なし

DocCountry :なし

### **PowerPoint 2013**

DocFormat : “PowerPoint 2013” “PowerPoint 2013 with Macro”  
 : “PowerPoint 2013 Slide Show”  
 : “PowerPoint 2013 Slide Show with Macro”  
 “PowerPoint 2013 Template”  
 ”PowerPoint 2013 Template with Macro"

DocSubFormat :なし

DocCountry :なし

**PowerPoint 2016/2019**

DocFormat : “PowerPoint 2016” “PowerPoint 2016 with Macro”  
 : “PowerPoint 2016 Slide Show”  
 : “PowerPoint 2016 Slide Show with Macro”  
 “PowerPoint 2016 Template”  
 “PowerPoint 2016 Template with Macro”  
 DocSubFormat : なし  
 DocCountry : なし

**RTF**

DocFormat : “Microsoft RTF”  
 DocSubFormat : なし  
 DocCountry : “Japanese” “English” “Simplified Chinese”  
 “Traditional Chinese” “Korean”

**Lotus1-2-3**

DocFormat : “Lotus1-2-3 V5” “Lotus1-2-3 97/98” “Lotus1-2-3 2000”  
 “Lotus1-2-3 Millennium Edition”  
 DocSubFormat : なし  
 DocCountry : “Japanese” “English”

**HTML**

DocFormat : “HTML”  
 DocSubFormat : なし  
 DocCountry : なし

**TEXT**

DocFormat : "TEXT"  
 DocSubFormat : "文字集合名"  
 DocCountry : なし

**AutoCAD (DXF)**

DocFormat : “AutoCAD GX-III DXF” “AutoCAD GX-5/R12 DXF”  
 “AutoCAD R13/LT95 DXF” “AutoCAD R14/LT97 DXF”  
 “AutoCAD 2000 DXF” “AutoCAD 2007 DXF” “AutoCAD 2010 DXF”  
 “AutoCAD GX-III DXF BIN” “AutoCAD GX-5/R12 DXF BIN”  
 “AutoCAD R13/LT95 DXF BIN” “AutoCAD R14/LT97 DXF BIN”

DocSubFormat : “AutoCAD 2000 DXF BIN”  
 DocCountry : なし

### **AutoCAD (DWG)**

DocFormat : “AutoCAD GX-III DWG” “AutoCAD GX-5/R12 DWG”  
 “AutoCAD R13/LT95 DWG” “AutoCAD R14/LT97 DWG”  
 “AutoCAD 2000 DWG” “AutoCAD 2007 DWG”  
 “AutoCAD 2010 DWG”  
 DocSubFormat : なし  
 DocCountry : “Japanese” “English” “Simplified Chinese”  
 “Traditional Chinese” “Korean”

### **PageMaker**

DocFormat : “Page Maker 6.0” “Page Maker 6.5/7.0”  
 DocSubFormat : “WIN\_V65/70” “MAC\_V65/70”  
 DocCountry : なし

### **ClarisWorks**

DocFormat : “Claris Works 4.0 WP” “Claris Works 4.0 SS”  
 “Claris Works 4.0 DB”  
 DocSubFormat : なし  
 DocCountry : “Japanese”

### **AppleWorks**

DocFormat : “Apple Works WP” “Apple Works SS” “Apple Works DB”  
 “Apple Works PT” “Apple Works DR” “Apple Works PR”  
 DocSubFormat : なし  
 DocCountry : なし

### **WordPerfect**

DocFormat : “WordPerfect 6.0” “WordPerfect 7/8/9”  
 DocSubFormat : なし  
 DocCountry : なし

### **Corel Presentations**

DocFormat : “Corel Presentations Slide Show 7/8/9”  
 DocSubFormat : なし  
 DocCountry : なし

### **Microsoft Works**

DocFormat : “MS Works 2000 WPS” “MS Works 2000 WDB”  
 “MS Works 2000 WKS”  
 DocSubFormat : なし  
 DocCountry : なし

### **Mac Write II**

DocFormat : “Mac Write II”  
 DocSubFormat : なし  
 DocCountry : なし

### **XML**

DocFormat : “XML” “Word XML” “Excel XML”  
 DocSubFormat : なし  
 DocCountry : なし

### **IGES**

DocFormat : “IGES”  
 DocSubFormat : なし  
 DocCountry : なし

### **QuarkXPress**

DocFormat : “QuarkXpress”  
 DocSubFormat : “3.3” “4”  
 DocCountry : なし

### **DocuWorks**

DocFormat : “DocuWorks v3” “DocuWorks v4/v5/v6” “DocuWorks v7”  
 DocSubFormat : なし  
 DocCountry : なし

### **Visio**

DocFormat : “Visio V4.0” “Visio V5.0” “Visio 2000/2002” “Visio 2003/2007/2010”  
 DocSubFormat :なし  
 DocCountry :なし

### Visio2013

DocFormat : “Visio 2013” “Visio 2013 with Macro”  
 “Visio 2013 Stencil” “Visio 2013 Stencil with Macro”  
 “Visio 2013 Template” “Visio 2013 Template with Macro”  
 DocSubFormat :なし  
 DocCountry :なし

### WPS Office

DocFormat : “KingSoft WPS 97” “KingSoft WPS 2000” “KingSoft WPS 2001”  
 “KingSoft WPS 2002” “KingSoft WPS 2003”  
 DocSubFormat :なし  
 DocCountry :なし

### LZH

DocFormat : "LZH"  
 DocSubFormat :なし  
 DocCountry :なし

### ZIP

DocFormat : "ZIP"  
 DocSubFormat :なし  
 DocCountry :なし

### RAR

DocFormat : "RAR"  
 DocSubFormat :なし  
 DocCountry :なし

### パスワード付き文書

DocFormat : "ファイルフォーマット"  
 DocSubFormat : "ファイルフォーマット補足情報"  
 DocCountry : "ドキュメント Country レコード情報"



返されるエラーコード

TextPorter のバージョンやパスワード付き文書の種類によって、

error code 2001 "CantDetectFile"

error code 3001 "ProtectedByPassword"

error code 3002 "InvalidFile"

詳しくは、

「5.7 パスワード付き PDF 文書のテキスト抽出」

「5.8 セキュリティ設定した PDF のテキスト抽出制御仕様」

「5.9 パスワード付き Microsoft Office, 一太郎のテキスト抽出」

を参照してください。

#### **OpenOffice Draw 拡張子: \*.sxd**

DocFormat : " OpenOffice Draw 1.0"

DocSubFormat : なし

DocCountry : "ドキュメント Country レコード情報"

返されるエラーコード

error code 3001 " ProtectedByPassword "

#### **OpenOffice Global 拡張子: \*.sxg**

DocFormat : " OpenOffice Global 1.0"

DocSubFormat : なし

DocCountry : "ドキュメント Country レコード情報"

返されるエラーコード

error code 3001 " ProtectedByPassword "

#### **OpenOffice Impress 拡張子: \*.sxi**

DocFormat : " OpenOffice Impress 1.0"

DocSubFormat : なし

DocCountry : "ドキュメント Country レコード情報"

返されるエラーコード

error code 3001 " ProtectedByPassword "

#### **OpenOffice Math 拡張子: \*.sxm**

DocFormat : " OpenOffice Math 1.0"

DocSubFormat : なし

DocCountry : なし

返されるエラーコード

error code 3001 " ProtectedByPassword "

**OpenOffice Writer 拡張子 : \*.sxw**

DocFormat : " OpenOffice Writer 1.0"

DocSubFormat : なし

DocCountry : "ドキュメント Country レコード情報"

返されるエラーコード

error code 3001 " ProtectedByPassword "

**OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4 Calc 拡張子 : \*.ods**

DocFormat : " OpenOffice 3.1 Calc"

DocSubFormat : なし

DocCountry : なし

**OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4 Presentation 拡張子 : \*.odp**

DocFormat : " OpenOffice 3.1 Presentation"

DocSubFormat : なし

DocCountry : なし

**OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4 Writer 拡張子 : \*.odt**

DocFormat : " OpenOffice 3.1 Writer"

DocSubFormat : なし

DocCountry : なし

**Outlook Express 拡張子 : \*.eml**

DocFormat : "EML"

DocSubFormat : なし

DocCountry : なし

**Outlook 拡張子 : \*.msg**

DocFormat : "MSG"

DocSubFormat : なし

DocCountry : なし

**Outlook 2007 拡張子 : \*.msg**

DocFormat : "MSG 2007"  
DocSubFormat : なし  
DocCountry : なし

**Outlook 2010 拡張子 : \*.msg**

DocFormat : "MSG 2010"  
DocSubFormat : なし  
DocCountry : なし

### 3.4.2 テキスト抽出関数

アプリケーションファイルからテキストを抽出します。

```
INT DMC_GetText_V5(Byte* Appfile, Byte* Txtfile, DMC_TEXTINFO_V5* TextInfo,
DMC_OLEERR_CALLBACK pFuncOleErr)
```

#### 引数

- Appfile : アプリケーションファイルパス名
- Txtfile : 出力テキストファイルパス名
- TextInfo : テキスト情報構造体。出力するテキストの詳細指定
- pFuncOleErr : コールバック関数のポインタ

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

#### 説明

- テキスト情報構造体 (DMC\_TEXTINFO\_V5)

```
typedef struct{
```

```
    Byte      GroupName[MAX_GROUP_NAME];
    Byte      DefLangName[MAX_LANG_NAME];
    BOOL      bBigEndian;
    DWord      Option;

    #define DMC_GETTEXT_OPT_KEISEN    0x00000001
    #define DMC_GETTEXT_OPT_TAG      0x00000002
    #define DMC_GETTEXT_OPT_RUBI    0x00000004
    #define DMC_GETTEXT_OPT_CRLF    0x00000008
    #define DMC_GETTEXT_OPT_CR      0x00000010
    #define DMC_GETTEXT_OPT_LF      0x00000020
    #define DMC_GETTEXT_OPT_U2028   0x00000040
    #define DMC_GETTEXT_OPT_U2029   0x00000080
    #define DMC_GETTEXT_OPT_SHEET   0x00000100
    #define DMC_GETTEXT_OPT_PWD     0x00000400
    #define DMC_GETTEXT_OPT_OLE     0x00001000
    #define DMC_GETTEXT_OPT_OLE1    0x00002000
    #define DMC_GETTEXT_OPT_OLE2    0x00004000
```

```

#define DMC_GETTEXT_OPT_OLE3      0x00008000
#define DMC_GETTEXT_OPT_OUT       0x00010000
#define DMC_GETTEXT_OPT_LOOP      0x00020000
#define DMC_GETTEXT_OPT_SHFTAG    0x00040000
#define DMC_GETTEXT_OPT_SHFHEAD   0x00080000
#define DMC_GETTEXT_OPT_SHEET1    0x00100000
#define DMC_GETTEXT_OPT_CELL      0x00200000
#define DMC_GETTEXT_OPT_SIZE      0x00400000
#define DMC_GETTEXT_OPT_PDFSYM    0x00800000
#define DMC_GETTEXT_OPT_CSV1      0x01000000
#define DMC_GETTEXT_OPT_CSV2      0x02000000
#define DMC_GETTEXT_OPT_ENDCODE   0x04000000
#define DMC_GETTEXT_OPT_NULL      0x08000000
#define DMC_GETTEXT_OPT_OWNERPWD1 0x10000000
#define DMC_GETTEXT_OPT_OWNERPWD2 0x20000000
#define DMC_GETTEXT_OPT_OWNERPWD3 0x40000000
#define DMC_GETTEXT_OPT_OWNERPWD4 0x80000000

```

#### DWord      Option1;

```

#define DMC_GETTEXT_OPT1_TEMP      0x00000001
#define DMC_GETTEXT_OPT1_INSERTF   0x00000002
#define DMC_GETTEXT_OPT1_INSERTF1  0x00000004
#define DMC_GETTEXT_OPT1_INSERTF2  0x00000008
#define DMC_GETTEXT_OPT1_INSERTF3  0x00000010
#define DMC_GETTEXT_OPT1_OWNERPWD5 0x00000020
#define DMC_GETTEXT_OPT1_COMPRESS   0x00000100
#define DMC_GETTEXT_OPT1_COMPRESS1 0x00000200
#define DMC_GETTEXT_OPT1_COMPRESS2 0x00000400
#define DMC_GETTEXT_OPT1_COMPRESS3 0x00000800
#define DMC_GETTEXT_OPT1_COMPRESS4 0x00001000
#define DMC_GETTEXT_OPT1_TRACK      0x00002000
#define DMC_GETTEXT_OPT1_COMPRESS5 0x00004000
#define DMC_GETTEXT_OPT1_INSERTF4  0x00008000
#define DMC_GETTEXT_OPT1_TXCONV     0x00010000
#define DMC_GETTEXT_OPT1_TXCONV2    0x00020000
#define DMC_GETTEXT_OPT1_OUTPUT_RAW_NL 0x00040000
#define DMC_GETTEXT_OPT1_QUOTE_QQ   0x00080000

```

```

Long          Size;
Word          Csv_c;
} DMC_TEXTINFO_V5, *LPDMC_TEXTINFO_V5;

```

#### **GroupName:変換先組み合わせ文字集合名称**

EUC-JP、EUC-JP-FIX、ISO-10646-UCS-2、ISO-10646-UCS-4、ISO-2022-JP、Shift\_JIS、UTF-16、UTF-8、WINDOWS31J、Shift\_JIS-2004、ISO-2022-JP-2004、EUC-JIS-2004、ISO\_8859-1、GB18030、GBK、Big5、KS\_C\_5601\_1987 のいずれかを指定します。

#### **DefLangName:変換元ファイルの言語指定**

Japanese、English、Simplified Chinese、Traditional Chinese、Korean のいずれかが指定できます。ファイル識別ライブラリで DocCountry が識別できるときは、DefLangName の指定がされていても DocCountry 情報を優先し適当な言語を DefLangName に設定します。DocCountry が識別できないときは外部から指定された言語を DefLangName にセットします。なにも指定されてない場合は「Japanese」に設定します。

#### **bBigEndian:エンディアン指定**

出力先テキストファイルの Endian を指定することができます。出力結果テキストファイルを読み込むアプリケーションにあわせて指定してください。

bBigEndian = 0 または DMC\_FALSE -> Little endian

bBigEndian = 1 または DMC\_TRUE -> big endian

#### **Option:オプション**

0 : オプションなし。

0 以外 : #define を参照して処理します。

オプションの指定内容は、後述「オプションの詳細説明」を参照

#### **Option 1:オプション 1**

0 : オプションなし。

0 以外 : #define を参照して処理します。

オプションの指定内容は、後述「オプション 1 の詳細説明」を参照

**Size:サイズ指定**

テキスト取りだし最大サイズ（バイト数）の指定。

**Csv\_c:データ区切り文字コード指定**

CSV のデータ区きり文字コードの指定。

指定できるコードは、0x09,0x0a,0x0d,0x20～0x7F のコードです。

■ **pFuncOleErr**

OLE 抽出でエラーになる場合、どの OLE の、どういうエラーなのかが分かるように構造体に記録してアプリケーションへ返すことができます。

この場合、DMC ライブラリ呼び出し側からエラーのコールバック関数を渡すようにします。

DMC ライブラリ側からは、エラーが発生したら、pFuncOleErr に LPDMC\_OLEERR を渡して、コールバックすることになります。

コールバック方式にしておけば、DMC ライブラリ呼び出し側で、エラーを無視したり、累積したり、自分の都合に合わせた処理ができます。

NULL を渡すと、コールバックしません。

**OLE エラー情報構造体の定義**

```
typedef struct {
    int      LevNum;          //エラーが発生した OLE の多重レベル(何段階ネストした OLE であるかを示します)
    Byte     DtctResult[256]; //識別結果
    DWord    ErrCode;         //エラーコード
} LDDMC_OLEERR;
```

**コールバック関数の定義**

```
typedef BOOL (*DMC_OLEERR_CALLBACK)
    (LDDMC_OLEERR* OleErr)
```

戻り値

0 を返せば、DMC ライブラリは処理続行します。

1 を返せば、処理打ち切ります。

## オプションの詳細説明

Option				説明
DMC_GETTEXT_OPT_KEISEN				全角文字罫線を出力します。 一太郎 6、OASYS のみ有効です。
DMC_GETTEXT_OPT_TAG				PowerPoints でスライドとノート を区別するタグを出力します。 (本仕様書の「6.3.2 抽出データ中 のタグ出力」を参照) DMC_GETTEXT_OPT_SHFTAG と同時指 定された場合は、無効となります。
DMC_GETTEXT_OPT_RUBI				ルビを抽出します。 Excel のみ有効。
DMC_GETTEXT_OPT_CRLF DMC_GETTEXT_OPT_CR DMC_GETTEXT_OPT_LF DMC_GETTEXT_OPT_U2028 DMC_GETTEXT_OPT_U2029				改行とパラグラフ分離マークを強 制的に指定のコードに置換しま す。
DMC_GETTEXT_OPT_SHEET				ページ抽出で各行の行頭にシート 名を付けます。Excel のみ有効。
DMC_GETTEXT_OPT_PWD				ユーザパスワード付き PDF ファイ ルや、パスワード付き(暗号化され た)Microsoft Office ファイルを 抽出します。
DMC_GETTEXT_OPT_OLE (PDF の場合、 DMC_GETTEXT_OPT1_IN SERTF 関連オプショ ンと同じ意味になり ます) (DMC_GETTEXT_OPT_OL E1 とは排他的なオプ ションであり、同時に 指定することはでき ません。指定したとき の動作は不定です)	1			OLE のテキストを抽出します。 一太郎のメインシート以外のシー トを抽出します。
	0	DMC_GETTEXT_OPT_OLE1	0	OLE を無視して抽出します。
			1	OLE が見つかったらエラーを返し ます。
	1	DMC_GETTEXT_OPT_OLE2	0	OLE のテキストを別ファイルに出 力します。
			1	OLE のテキストを本文テキストと 同じファイルに出力します。
		DMC_GETTEXT_OPT_OLE3	0	OLE の抽出が失敗した時、エラーを 返して、このファイルのテキスト 抽出を中止します。
			1	OLE の抽出が失敗した時、この OLE のテキスト抽出を中止して次の OLE あるいは本文テキストの抽出 を続行します。



DMC_GETTEXT_OPT_OUT			<p>ログファイルを出力します。</p> <p>★Log ファイル名 テキスト抽出先ファイル名 + Engine 名 + .log 例: 抽出元 abc.xls のログファイルは abcxls.log です。</p> <p>★Log ファイルは抽出先ファイルと同じフォルダに作られます。 ストリーム出力の場合は、抽出元ファイルと同じフォルダに作られます。</p>
DMC_GETTEXT_OPT_LOOP			無限ループ自動検知をします。
DMC_GETTEXT_OPT_SHFTAG			<p>Word/Excel/PowerPoint/PDF のヘッダー／フッター、及び Excel のシート名を特別な形式で出力する。</p> <p>1. Excel のシート名 〈SheetN:Sheet 名〉</p> <p>2. Excel のヘッダー／フッター 〈SheetN_Header:Header 内容〉 〈SheetN_Footer:Footer 内容〉 (N:Sheet No.)</p> <p>3. Word 〈Header:Header 内容〉 〈Footer:Footer 内容〉</p> <p>4. PowerPoint 〈SlideFooter:Footer 内容〉 〈NotesHeader:Header 内容〉 〈NotesFooter:Footer 内容〉</p> <p>5. PowerPoint /PDF ページの切れ目 (先頭は除く) 〈page〉を入れる</p>
DMC_GETTEXT_OPT_SHFHEAD			Word/Excel/PowerPoint のヘッダー／フッターを出力しません。
DMC_GETTEXT_OPT_SHEET1			Excel、Lotus のシート名を出力しません。
DMC_GETTEXT_OPT_CELL			Excel、Lotus の空白セルを出力しません。
DMC_GETTEXT_OPT_SIZE			<p>抽出結果テキストファイルの Max サイズを指定します。例: 1 MB を指定した場合、抽出処理を 1 MB まで終了させます。</p> <p>★ファイルの最後の文字 (2 バイト以上) が 1 文字として抽出できない場合、切り捨てます。</p> <p>★OLE 抽出の設定で、別ファイルに出力、メインファイルと同じフ</p>

			ファイルに出力するかに関わらず、指定されたファイルサイズで抽出します。
DMC_GETTEXT_OPT_PDFSYM			PDF の Symble 文字を抽出しません。
DMC_GETTEXT_OPT_CSV1			表計算の文字列データの両側は「」で括りません。
DMC_GETTEXT_OPT_CSV2		0	表計算はカンマ (2ch)、AutoCAD はデータ区切りはありません。
		1	外部から指定。 例: TAB (09h)、Space (20h) ★表計算、AutoCAD のみ有効
DMC_GETTEXT_OPT_ENDCODE			Word、PowerPoint、HTML は改行コードの出力前に NULL を挿入します (0x00+改行コード)
DMC_GETTEXT_OPT_OWNERPWD1 DMC_GETTEXT_OPT_OWNERPWD2 DMC_GETTEXT_OPT_OWNERPWD3 DMC_GETTEXT_OPT_OWNERPWD4			セキュリティ設定された PDF ファイルのテキスト抽出を制御します。 詳細は、付録「セキュリティ設定したPDFのテキスト抽出制御仕様」を参照してください。

## オプション 1 の詳細説明

Option1		説明		
DMC_GETTEXT_OPT1_TEMP				テンポラリーファイルのフォルダをシステムのデフォルトのフォルダから、抽出先ファイルのフォルダに変更します。 ストリーム出力の場合は、抽出元ファイルのフォルダに変更します。 変更先のフォルダが、書き込み禁止の場合、動作は不定です。
DMC_GETTEXT_OPT1_INSERTF (DocuWorks, eml, msg PDF の場合に有効) (DocuWorks, PDF の場合、DMC_GETTEXT_OPT1_OLE 関連オプションと同じ意味になります) (DMC_GETTEXT_OPT1_INSERTF1 とは排他的なオプションであり、同時に指定することはできません。指定したときの動作は不定です)	1			添付ファイル及びOLE 組み込みされたファイルのテキストを抽出します。
	0	DMC_GETTEXT_OPT1_INSERTF1	0	添付ファイル及びOLE 組み込みされたファイルが無視します。
			1	添付ファイル及びOLE 組み込みされたファイルが見つかった場合エラーを返します。
	1	DMC_GETTEXT_OPT1_INSERTF2	0	添付ファイル及びOLE 組み込みされたファイルのテキストを別ファイルに出力します。
			1	添付ファイル及びOLE 組み込みされたファイルのテキストを本文テキストと同じファイルに出力します。 DMC_GETTEXT_OPT1_COMPRESS2 を同時指定してください。
		DMC_GETTEXT_OPT1_INSERTF3	0	添付ファイル及びOLE 組み込みされたファイルの抽出が失敗した場合、エラーを返して、このファイルのテキスト抽出を中止します。
			1	添付ファイル及びOLE 組み込みされたファイルの抽出が失敗した場合、この添付ファイル及びOLE 組み込みされたファイルの抽出を中止して、次の添付ファイル及びOLE 組み込みされたファイルあるいは本文テキストの抽出を続行します。
		DMC_GETTEXT_OPT1_INSERTF4	0	INSERTF2 の指定時、ファイル情報 (「attachmemt_name:」) を出力する。
			1	INSERTF2 の指定時、ファイル情報 (「attachmemt_name:」) を出力しない。
DMC_GETTEXT_OPT1_OWNERPWD5				セキュリティ設定されたPDF ファイルのテキスト抽出を制御します。 詳細は、付録「セキュリティ設定したPDF のテキスト抽出制御仕様」を参照してください。

DMC_GETTEXT_OPT1_COMPRESS (LZH、ZIP、RAR のみ有効) (DMC_GETTEXT_OPT1_COMPRESS1 とは排他的なオプションであり、同時に指定することはできません。指定したときの動作は不定です)	1			圧縮ファイルのテキストを抽出します。
	0	DMC_GETTEXT_OPT1_COMPRESS1	0 1	圧縮ファイルを無視します。 圧縮ファイルが見つかった場合エラーを返します。
	1	DMC_GETTEXT_OPT1_COMPRESS2	0	圧縮ファイルから抽出されたテキストをファイル毎に別ファイルに出力します。 <sup>1</sup>
			1	圧縮ファイルから抽出されたテキストを同じファイルに出力します。
		DMC_GETTEXT_OPT1_COMPRESS3	0	圧縮ファイルの抽出が失敗した場合、エラーを返して、圧縮ファイルのテキスト抽出を中止します。
			1	圧縮ファイル中のファイルのテキスト抽出が失敗した場合、該当ファイルの抽出を中止して、次のファイルあるいは本文テキストの抽出を続行します。
		DMC_GETTEXT_OPT1_COMPRESS4	0 1	圧縮ファイルのフォルダ構成を保持します。 圧縮ファイルのフォルダ構成を無視して、すべてのテキストを指定した抽出先フォルダに出力します。
DMC_GETTEXT_OPT1_COMPRESS5			0	COMPRESS2 の指定時、ファイル情報 (「Filename:」「Content:」) を出力する。
			1	COMPRESS2 の指定時、ファイル情報 (「Filename:」「Content:」) を出力しない。
DMC_GETTEXT_OPT1_TRACK (Word, RTF, WPS のみ有効)	1			文書の変更履歴記録を抽出しない。
DMC_GETTEXT_OPT1_TXCONV	1			テキストをコード変換する。
DMC_GETTEXT_OPT1_TXCONV2	1			テキストをコード変換する際に元のエンコードが判別できない場合は、テキストを書き出さない。

<sup>1</sup> ストリーム抽出時は無効

DMC_GETTEXT_OPT1_OUTPUT_RAW_NL (このオプションを指定すると、DMC_GETTEXT_OPT_CSV1 は無視され、必ず、「”」で括ります)				Excel で、セルやヘッダなどの内部に含まれる改行を空白にせず、改行を出力する。出力する改行コードは、DMC_GETTEXT_OPT_CRLF (デフォルト) DMC_GETTEXT_OPT_CR DMC_GETTEXT_OPT_LF DMC_GETTEXT_OPT_U2028 DMC_GETTEXT_OPT_U2029 の指定に従う。
DMC_GETTEXT_OPT1_QUOTE_QQ (DMC_GETTEXT_OPT1_OUTPUT_RAW_NLと関連したオプション)				Excel で、「”」でテキストを括るとき、テキスト中の「”」を「¥”」と出力するのではなく、「"""」と出力する。

### 3.4.3 テキスト抽出関数（ストリーム出力）

アプリケーションファイルからテキストをストリームに抽出します。

**INT DMC\_GetTextStream\_V5(Byte\* Appfile, ostream\* pOutputStream, TEXTINFO\_V5\* TextInfo, DMC\_OLEERR\_CALLBACK pFuncOleErr)**

#### 引数

- Appfile : アプリケーションファイルパス名
- pOutputStream : 出力するストリーム
- TextInfo : 出力するテキストの詳細指定（詳細は「3.4.2 テキスト抽出関数」を参照）
- pFuncOleErr : コールバック関数のポインタ（詳細は「3.4.2 テキスト抽出関数」を参照）

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

### 3.4.4 プロパティ抽出仕様

指定されたファイルから「プロパティ」情報を抽出し、プロパティ構造体に設定します。

**INT DMC\_GetProperty\_V5(Byte\* Appfile, DMC\_TEXTINFO\_V5\* TextInfo, DMC\_PROPERTY\* Property)**

#### 引数

- Appfile : アプリケーションファイル名
- TextInfo : 出力するプロパティの詳細指定（詳細は「3.4.2 テキスト抽出関数」を参照）
- Property : プロパティ構造体アドレス

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

#### 説明

- プロパティ構造体

```
typedef struct {
    char title[MAXCHARBUF];
    char author[MAXCHARBUF];
    char keyword[MAXCHARBUF];
    char subject[MAXCHARBUF];
    char comment[MAXCHARBUF];
    char manager[MAXCHARBUF];
    char company[MAXCHARBUF];
    char category[MAXCHARBUF];
    char createdate[MAXDATABUF];
    char revision[MAXCHARBUF];
    char lastrevisor[MAXCHARBUF];
    char revisioncount[MAXDATABUF];
    char lastprintdate[MAXDATABUF];
    char edittime[MAXDATABUF];
    char creator[MAXCHARBUF];
    char producer[MAXCHARBUF];
    char encryptionflag[32];
}
```

```
char slides[MAXDATABUF];
char paragraphs[MAXDATABUF];
char bytes[MAXDATABUF];
char notes[MAXDATABUF];
char presentation[MAXCHARBUF];
char doctype[MAXDATABUF];
char lastsavetime[MAXDATABUF];
char owner[MAXDATABUF];
char abstract[MAXDATABUF];
char account[MAXDATABUF];
char address[MAXDATABUF];
char attachments[MAXDATABUF];
char authorization[MAXDATABUF];
char bill_to[MAXDATABUF];
char blind_copy[MAXDATABUF];
char carbon_copy[MAXDATABUF];
char checked_by[MAXDATABUF];
char client[MAXDATABUF];
char department[MAXDATABUF];
char descriptive_name [MAXDATABUF];
char descriptive_type[MAXDATABUF];
char destination[MAXDATABUF];
char disposition[MAXDATABUF];
char division[MAXDATABUF];
char document_number[MAXDATABUF];
char editor[MAXDATABUF];
char forward_to[MAXDATABUF];
char group[MAXDATABUF];
char language[MAXDATABUF];
char mail_stop[MAXDATABUF];
char matter[MAXDATABUF];
char office[MAXDATABUF];
char project[MAXDATABUF];
char publisher[MAXDATABUF];
char purpose[MAXDATABUF];
char received_from[MAXDATABUF];
```



```

char recorded_by[MAXDATABUF];
char recorded_date[MAXDATABUF];
char reference[MAXDATABUF];
char revision_date[MAXDATABUF];
char revision_notes[MAXDATABUF];
char section[MAXDATABUF];
char security[MAXDATABUF];
char source[MAXDATABUF];
char status[MAXDATABUF];
char telephone_number[MAXDATABUF];
char typist[MAXDATABUF];
char version_date[MAXDATABUF];
char version_notes[MAXDATABUF];
char version_number[MAXDATABUF];
} DMC_PROPERTY;

```

#### ■ 構造体の詳細

	説明
title	タイトル
author	作者
keyword	キーワード
subject	サブジェクト
comment	コメント
manager	管理者
company	会社
category	分類
createdate	ファイル作成日付
revision	修正説明(Lotus1-2-3)
lastrevisor	最後の修正者
revisioncount	修正回数
lastprintdate	最後のプリント日付(Lotus1-2-3)
edittime	編集時間
creator	作成者(PDF)
producer	変換(PDF)
encryptionflag	暗号
slides	スライド枚数(PowerPoint)
paragraphs	段落数(PowerPoint)
bytes	バイト数(PowerPoint)
notes	ノートの枚数(PowerPoint)
presentation	Slide 表現様式(PowerPoint)
doctype	ファイル種類(OASYS/Win)

lastsavetime	最終セーブ日付
owner	オーナー(OASYS/Win : 結合形式)
abstract	摘要 (以下 Wordperfect)
account	勘定
address	アドレス
attachments	アクセサリー
authorization	オーソリゼーション
bill_to	小切手
blind_copy	ブラインドコピー
carbon_copy	複本
checked_by	照合
client	クライアント
department	部門
descriptive_name	記述名
descriptive_type	記述タイプ
destination	目的
disposition	配置
division	部分
document_number	文書数
editor	編集者
forward_to	フォワード
group	グループ
language	言語
mail_stop	メール停止
matter	問題
office	オフィス
project	プロジェクト
publisher	出版者
purpose	用途
received_from	受信先
recorded_by	記録
recorded_date	記録日付
reference	参照
revision_date	改訂日付
revision_notes	改訂注記
section	セクション
security	セキュリティ
source	ソース
status	状態
telephone_number	電話番号
typist	入力者
version_date	ファイル更新日
version_notes	バージョン注記
version_number	バージョン数

- プロパティ抽出構造体は各ファイルフォーマット共通です。

- プロパティの情報は、各ファイルフォーマットによって抽出できる項目が異なります。プロパティ情報として保持していない項目は Null に設定します。
- 構造体の説明に ( ) が有る場合は、( ) 内に記載した各ファイルフォーマットの固有情報を表します。( ) 内に指定したファイルフォーマットのプロパティ抽出で無い限り、その項目には情報をセットしません。
- ファイル作成日時等、日付情報はデータに格納されたまま出力します。MiciroSoft Office 等、日付データを GMT 形式で格納しているアプリケーションデータから抽出した場合は、アプリケーションで開いたときに表示される情報と、プロパティ抽出された情報等で誤差が出る場合があります。

各エンジンの抽出仕様（Y：抽出できるプロパティ、空は抽出できないプロパティ）

	2007 2010 2013 2016 2019	xls	doc	ppt	rtf	pdf	oas	oa2	vsd  visio 2013	123	wk4	jaw jbw	jfw	jtd
title	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
author	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y		Y	Y
keyword	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y		Y	Y
subject	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y			
comment	Y	Y	Y	Y	Y			Y	Y	Y	Y			Y
manager	Y	Y	Y	Y	Y				Y					
company	Y	Y	Y	Y					Y					
category	Y	Y	Y	Y	Y				Y					
createdate		Y			Y	Y	Y	Y			Y		Y	Y
revision										Y	Y			
lastrevisor				Y	Y					Y	Y		Y	Y
revisioncount				Y						Y	Y		Y	Y
lastprintdate										Y				
edittime										Y	Y			
creator						Y								
producer						Y								
encryptionflag						Y								
slides	Y*1			Y										
paragraphs	Y*2			Y										
bytes				Y										
notes				Y										
presentation				Y										
doctype							Y	Y						
lastsavetime							Y	Y			Y	Y		
owner														
abstract														
account														
address														
attachments														
authorization														
bill_to														
blind_copy														
carbon_copy														
checked_by														
client														
department														
descriptive_name														
descriptive_type														
destination														
disposition														
devision														
document_number														
editor														
forward_to														
group														
language														

mail_stop														
matter														
office														
project														
publisher														
purpose														
received_from														
recorded_by														
recorded_data														
reference														
revision_date														
revision_notes														
section														
security														
source														
status	Y													
telephone_number														
typist														
version_date														
version_notes														
version_number														

- ◆ Office2019 Microsoft Word2019/Excel 2019/PowerPoint 2019  
\*1 PowerPoint 2019 のみ \*2 Word 2019,PowerPoint 2019 のみ
- ◆ Office2016 Microsoft Word2016/Excel 2016/PowerPoint 2016  
\*1 PowerPoint 2016 のみ \*2 Word 2016,PowerPoint 2016 のみ
- ◆ Office2013 Microsoft Word2013/Excel 2013/PowerPoint 2013  
\*1 PowerPoint 2013 のみ \*2 Word 2013,PowerPoint 2013 のみ
- ◆ Office2010 Microsoft Word2010/Excel 2010/PowerPoint 2010  
\*1 PowerPoint 2010 のみ \*2 Word 2010,PowerPoint 2010 のみ
- ◆ Office2007 Microsoft Word2007/Excel 2007/PowerPoint 2007  
\*1 PowerPoint 2007 のみ \*2 Word 2007,PowerPoint 2007 のみ
- ◆ xls Microsoft Excel V4/V5/V7(95)/97/2000/XP/2003  
Microsoft Excel 98 for Macintosh/2001 for Macintosh
- ◆ doc Microsoft Word V6/V7(95)/97/98/2000/XP/2003  
Microsoft Word 2001 for Macintosh
- ◆ ppt Microsoft PowerPoint 95/97/2000/XP/2003  
Microsoft PowerPoint 2001 for Macintosh
- ◆ rtf Microsoft RTF 1.3/1.4/1.5
- ◆ pdf 1.2/1.3/1.4/1.5/1.6/1.7
- ◆ oas 富士通 OASYS オンライン形式
- ◆ oa2 富士通 OASYS Win oa2/oa3/分離形式
- ◆ vsd Visio V4/V5/2000/2002/2003/2007/2010

- ◆ 123 Lotus 1-2-3 97/98/MillenniumEdition9.5/2000/OASYS 1-2-3 V6/V7/V8
- ◆ wk4 Lotus 1-2-3 R5
- ◆ jaw JUSTSYSTEM 一太郎 V5
- ◆ jbw JUSTSYSTEM 一太郎 V6/V6.3/dash2
- ◆ jfw JUSTSYSTEM 一太郎 7
- ◆ jtd JUSTSYSTEM 一太郎 Lite、8 から 13、2004 から 2019

	dxf	dwg	cwj	pm p65	wdb	wpd	wps	awj	shw xdw	igs	xml (*1)	xml (*2)	wpso
title	Y	Y	Y	Y	Y		Y	Y	Y		Y	Y	Y
author	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
keyword	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	
subject	Y	Y		Y	Y	Y	Y		Y		Y	Y	
comment	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
manager				Y	Y						Y	Y	
company				Y	Y					Y	Y	Y	
category			Y	Y	Y	Y	Y	Y			Y	Y	
createdate	Y	Y				Y					Y	Y	
revision			Y					Y			Y		
lastrevisor	Y	Y							Y		Y	Y	
revisioncount						Y			Y				
lastprintdate													
edittime	Y	Y									Y		
creator													
producer													
encryptionflag													
slides													
paragraphs											Y		
bytes													
notes													
presentation													
doctype													
lastsavetime		Y				Y					Y	Y	
owner						Y							
abstract						Y					Y		
account						Y							
address						Y					Y	Y	
attachments						Y							
authorization						Y							
bill_to						Y							
blind_copy						Y							
carbon_copy						Y							
checked_by						Y					Y		
client						Y					Y		
department						Y							
descriptive_name						Y							
descriptive_type						Y							
destination						Y							
disposition						Y							
devision						Y							
document_number						Y							
editor						Y							
forward_to						Y							
group						Y							
language						Y							
mail_stop						Y							
matter						Y					Y		
office						Y							

project						Y							
publisher						Y							
purpose						Y							
received_from						Y							
recorded_by						Y							
recorded_data						Y							
reference						Y							
revision_date	Y												
revision_notes						Y							
section						Y							
security						Y							
source						Y							
status						Y							
telephone_number						Y							
typist						Y					Y		
version_date						Y							
version_notes						Y							
version_number						Y					Y	Y	

- ◆ dxf AutoCAD R14 DXF
- ◆ dwg AutoCAD R14 DWG
- ◆ cwj ClarisWorks 4.0
- ◆ pm Adobe PageMaker 6.0/6.5
- ◆ wdb WordPerfect6.0/7/8/9
- ◆ wpd Microsoft Works 2000
- ◆ wps Microsoft Works 2000
- ◆ awj Apple Works 6
- ◆ igs IGES
- ◆ shw Corel Presentations 9
- ◆ xdw DocuWorks v4/v5/v6/v7/v8
- ◆ xml(\*1) Word XML
- ◆ xml(\*2) Excel XML
- ◆ wpso WPS Office 97/2000/2002/2003

★html、Ms Works 2000、Mac Write II、QuarkXpress3.3/4、eml、msg、テキストの  
 ロパティ抽出機能はありません。



### 3.4.5 頁抽出関数

アプリケーションファイルから指定された頁のテキストを抽出します。

```
INT DMC_GetPageText_V5(Byte* Appfile, Byte* Txtfile, DMC_TEXTINFO_V5* TextInfo,
    int* Pages, DMC_OLEERR_CALLBACK pFuncOleErr)
```

#### 引数

- Appfile : アプリケーションファイルパス名
- Txtfile : 出力テキストファイルパス名
- TextInfo : 出力するテキストの詳細指定（詳細は「3.4.2 テキスト抽出関数」を参照）
- Pages : 0 指定時、対象ファイルの総頁数を取得します  
           >0 指定時、指定した頁のテキストを抽出します
- pFuncOleErr : コールバック関数のポインタ（詳細は「3.4.2 テキスト抽出関数」を参照）

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

#### 説明

下記のファイルフォーマットは、頁抽出が可能です

Microsoft Excel V5/V7(95)/97/2000/XP/2003/2007/2010/  
 2013/2016/2019

Microsoft Excel 98 for Macintosh

Microsoft Excel 2001 for Mac

Lotus1-2-3 R5

Lotus1-2-3 97/98/2000

Microsoft PowerPoint 95/97/2000/XP/2003/2007/2010/  
 2013/2016/2019

Microsoft PowerPoint 2001 for Macintosh

PDF 1.2/1.3/1.4/1.5/1.6/1.7

Adobe PageMaker6/6.5/7.0

DocuWorks v4/v5/v6/v7/v8

Visio v4/v5./2000/2002/2003/2007/2010/2013

ページは、印刷上のページとはならない場合があります。たとえば、Excel  
 ではシート、PowerPoint ではスライドをページとみなします。

上記以外のファイルフォーマットでは、動作は不定です。

### 3.4.6 頁抽出関数（ストリーム出力）

アプリケーションファイルから指定された頁のテキストをストリームへ抽出します。

```
INT DMC_GetPageTextStream_V5(Byte* Appfile, ostream* pOutputStream, DMC_TEXTINFO_V5*  
TextInfo, int* Pages, DMC_OLEERR_CALLBACK pFuncOLEErr)
```

#### 引数

- Appfile : アプリケーションファイルパス名
- pOutputStream : 出力するストリーム
- TextInfo : 出力するテキストの詳細指定（詳細は「3.4.2 テキスト抽出関数」を参照）
- Pages : 0 指定時、対象ファイルの総頁数を取得します。  
>0 指定時、指定した頁のテキストを抽出します。
- pFuncOLEErr : コールバック関数のポインタ（詳細は「3.4.2 テキスト抽出関数」を参照）

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

#### 説明

頁抽出が可能なフォーマットについては、「DMC\_GetPageText\_V5」の説明を参照してください。

### 3.4.7 パスワード付きファイルのテキスト抽出関数

ユーザパスワード付き PDF ファイルや、パスワード付き(暗号化された)Microsoft Office ファイルから、テキストを抽出します。

```
INT DMC_GetPwdText_V5(Byte* Appfile, Byte* Txtfile, DMC_TEXTINFO_V5* TextInfo,
Byte* Password, DMC_OLEERR_CALLBACK pFuncOleErr)
```

#### 引数

- Appfile、Txtfile、TextInfo、pFuncOleErr（詳細は「3.4.2 テキスト抽出関数」を参照）
- Password : パスワード

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

### 3.4.8 パスワード付きファイルの頁抽出関数

ユーザパスワード付きファイルや、パスワード付き(暗号化された)Microsoft Office ファイルから、指定した頁のテキストを抽出します。

```
INT DMC_GetPwdPageText_V5(Byte* Appfile, Byte* Txtfile, DMC_TEXTINFO_V5* TextInfo,
int* Pages, Byte* Password, DMC_OLEERR_CALLBACK pFuncOleErr)
```

#### 引数

- Appfile、Txtfile、TextInfo、Pages、pFuncOleErr（詳細は「3.4.5 頁抽出関数」を参照）
- Password : パスワード

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

### 3.4.9 パスワード付きファイルのプロパティ抽出関数

ユーザパスワード付き PDF ファイルや、パスワード付き(暗号化された)Microsoft Office ファイルから、プロパティを抽出します。

**INT DMC\_GetPwdProperty\_V5 (Byte\* Appfile, LPDMC\_TEXTINFO\_V5 TextInfo,  
DMC\_PROPERTY\* Property, Byte\* Password)**

#### 引数

- Appfile、TextInfo、Property (詳細は「3.4.4 プロパティ抽出関数」を参照)
- Password : パスワード

#### 戻り値

- 成功は、0 を返します
- 失敗は、エラー番号を返します

## 4. テキスト抽出仕様

---

### 4.1 共通仕様

#### 4.1.1 制御コード

- ワードプロ本文中の制御コードのうち、改行コード以外の制御コードは削除します。

#### 4.1.2 定義外文字

- 抽出先の符号化方式で使われる基本文字集合にない文字は類似の文字(1 文字または 1 文字の組合せ)にマップします。
- 類似の文字が無い場合は、"=" (2 バイト)、"?" (1 バイト)に書き換えて出力します。

#### 4.1.3 ユーザ外字

- 抽出先の符号化文字集合で、ユーザ定義文字を使用できない場合は、アプリケーションのユーザ外字は、"=" (2 バイト)、"?" (1 バイト)に書き換えて出力します。
- 抽出先で指定した文字集合に合わせてコードを変換します。例えば抽出元データが Shift-JIS で、ファイルの中にユーザー外字領域 F040 に割り当てられた文字が含まれていた場合、抽出先に UTF-8 を選択した上で処理を行うと Shift-JIS F040 が Unicode の 0xE001 に変換されて出力されます。  
このため、Unicode の 0xE001 に文字が登録されていればテキストは表示可能となります。

#### 4.1.4 OLE オブジェクト抽出

- 本ライブラリは、OLE オブジェクトの 3 階層まで抽出できます。3 階層以上の場合、エラーコード 4005 を返します。
- 本ライブラリは、リンクされるオブジェクトは抽出できません。
- 次の表に、主なアプリケーションについて、OLE の対応を示します。表の列は、OLE の親文書。行は、OLE で埋め込まれる文書を示します。たとえば、列が Excel 2003 で、行が Word 2003 の場所には○が付いているので、Excel 2003 に Word 2003 を埋め込んだ場合は、抽出可能となります。
- 暗号化された Microsoft Office ファイルが、その中に OLE で埋め込まれたものを含む場合、OLE で埋め込まれたものは、抽出できません。

## 主要アプリケーションの OLE 対応表

	Word 2003	Excel 2003	PowerPoint 2003	Word 2007/2010/2013 /2016/2019	Excel 2007/2010/2013 /2016/2019	PowerPoint 2007/2010/2013 /2016/2019	PDF	一太郎 2004-2019
Word 2003	○	○	○	○	○	○	○	○
Excel 2003	○	○	○	○	○	○	○	○
PowerPoint 2003	○	○	○	○	○	○	○	○
Word 2007/2010/2013 /2016/2019	○	○	○	○	○	○	○	○
Excel 2007/2010/2013 /2016/2019	○	○	○	○	○	○	○	○
PowerPoint 2007/2010/2013 /2016/2019	○	○	○	○	○	○	○	○
PDF	○	○	○	○	○	○	○	○
一太郎 2004-2019	○	○	○	○	○	○	○	×
OASYS	×	○	×	×	×	×	○	×
DocuWorks	×	×	×	×	×	×	○	×
OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4 Writer	×	×	×	×	×	×	○	×
OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Calc	×	×	×	×	×	×	○	×
OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Presentation	×	×	×	×	×	×	○	×

	OA SYS	DocuWorks	OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Writer	OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Calc	OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Presentation
Word 2003	×	○	×	×	×
Excel 2003	×	○	×	×	×
PowerPoint 2003	×	○	×	×	×
Word 2007/2010/2013//20 16/2019	×	○	×	×	×
Excel 2007/2010/2013/20 16/2019	×	○	×	×	×
PowerPoint 2007/2010/2013/20 16/2019	×	○	×	×	×
PDF	×	○	×	×	×
一太郎 2004-2019	×	○	×	×	×
OASYS	×	○	×	×	×
DocuWorks	×	○	×	×	×
OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Writer	×	×	×	×	×
OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Calc	×	×	×	×	×
OpenOffice.org 3.1/3.2/3.3 Libre Office 3.3/3.4 Presentation	×	×	×	×	×



#### 4.1.5 圧縮ファイルからの抽出

- パスワードで保護したファイルからテキスト抽出はできません。
- OASYS 分離形式ファイル (\*.doc, \*.fmt) のテキスト抽出には、圧縮ファイル中のファイルを一つづつ解凍→テキスト抽出→解凍されたファイルを削除という処理をしている為に、テキスト抽出はできません。
- 圧縮ファイル内のファイルに対してプロパティ抽出、ページ抽出、PDF のパスワード文書の抽出には対応していません。
- LZH 形式は、LHa5/LHa6/LHa7 をサポートしております。
- 自己解凍形式の圧縮ファイルには対応していません。

#### 4.1.6 ストリームへの抽出

- ストリームからのテキスト抽出はできません。
- パスワードを要求する PDF や Microsoft Office ファイルのストリームへの抽出はできません。
- Windows x64 版でストリームを使用する場合は、`setlocale()`にてロケールの設定を行ってください。

#### 【制限事項】

抽出されたテキストの順番は必ずしもレイアウトの表示結果と一致しません。

## 4.2 ワープロ文書

### 4.2.1 全角文字罫線（一太郎、OASYS、OASYS オンラインのみ有効）

- アプリケーションが全角文字罫線（罫線の高さが 1 行を占め、幅が全角 1 文字分を占める罫線）を使用している場合、全角文字罫線を出力する/しないを切りかえることができます。
- 全角文字罫線を出力する場合は、文字罫線コードに置き換えて出力します。
- 全角文字罫線を出力しない場合は、全角空白コードに置き換えて出力します。

### 4.2.2 その他の罫線

- 全角文字罫線以外の種類の罫線は、削除します。

### 4.2.3 表

- 表は解除し、セルの内容を文字列として抽出します。

#### 4.2.4 RTF

- 箇条書き

箇条書きの行頭文字は、抽出先文字集合はどれが指定されても「??」を出力します。

- 段落番号とアウトライン

段落番号とアウトライン番号は、指定された抽出先文字集合に該当文字がある場合は抽出できますが、ない場合は、4.1.2の「定義外文字」の仕様に従って出力します。

- 図形、イメージ、線画、枠、数式は無視します。

- Word、RTF 文書に挿入された自動更新の日付、時間は正しく抽出できません。

- Word、RTF 文書の特殊文字は一部抽出できません。

- Word、RTF ファイルのフィールドの内容は一部抽出できません。

#### 4.2.5 一太郎 8 から一太郎 13、一太郎 2004 から一太郎 2019

- ファイルの識別結果としては、「Ichitaro Document」となります。
- 圧縮して保存した文書は識別、抽出できません。
- 文書中の特定の行に付ける行番号は抽出できません。
- マスキング

フルード、入力ガイド内の文字は埋め文字で指定した文字に置き換わる場合、その指定した文字の抽出はできますが、マスキング文書ではレイアウト枠が塗りつぶし色で設定した色で塗りつぶされる場合、枠内の文字の抽出はできません。

### 4.3 プレゼンテーションファイル抽出仕様

#### 4.3.1 テキスト抽出処理概要

プレゼンテーション・ファイルからは、スライドとノートのテキストを抽出します。  
(スライド番号は抽出しません。)

#### 4.3.2 抽出データ中のタグ出力

オプションで「タグを出力する」を指定した場合、抽出時に以下のタグを付加して出力します。

- <slide>、</slide>、<notes>、</notes>などのタグを出力します。
- タグの出力仕様

スライド 1 : <slide></slide>

スライド 2 : <slide></slide>

・

・

・

スライド n : <slide></slide>

ノート 1 : <notes></notes>

ノート 2 : <notes></notes>

・

・

・

ノート n : <notes></notes>

※スライド毎に<slide></slide>でスライドからの抽出データを括り、ノートからの抽出データを<slide>外に順番に出力します。

## 4.4 表計算

- CSV 形式でテキストファイルに出力します。
- 行
  - ① ワークシートの一行を文字列の一行として出力します。
  - ② 行は上から順に出力します。
  - ③ 一行の終了には改行コードを出力します。
  - ④ データが存在しない行は改行コードのみ出力します。
- 列
  - ① 一行内の出力は、列の先頭から列順に出力します。
  - ② 列間は「,」で区切ります。
  - ③ データの無いセルは、データ無しとして出力します。この場合「前セルデータ, 後セルデータ」といった形で、「列区切りのカンマ」が連続して出力されることになります。但し、データが後ろに続かない場合は、最後の「,」は出力しません。
- セル

文字データセルを「"」で括って出力する場合、出力する文字列に「"」が含まれる場合は「\」でエスケープして「\"」と出力します。

DMC\_GETTEXT\_OPT1\_QUOTE\_QQ を指定すると、「"""」と出力します。

「'」で括らずに出力する場合、出力する文字列に「'」が含まれる場合は、「\」でエスケープしません。

- ① フォント、配置、罫線、パターン書式は全て無視します。
- ② 色属性以外の表示書式は、反映して出力します。表示書式を反映した場合に、数値データセルであるにも関わらず数値文字でない文字が含まれてしまう場合は、文字列として出力します。（指数表現など）
- ③ 数値文字は「0～9」の数値と符号である「+」「-」と小数点「.」から構成される文字列です。符号は先頭になければなりません。また小数点は数値文字列中に一つしか存在してはなりません。
- ④ 数値は 10 進数で表現されるものとします。
- ⑤ Excel2007 からの抽出時、小数点以下の数値が丸められて抽出される場合があります。
- ⑥ セル内改行コードは半角空白に出力し、行を連結して出力します。

DMC\_GETTEXT\_OPT1\_OUTPUT\_RAW\_NL を指定すると、セル内改行を半角空白にせず、出力します。出力する改行コードは、

DMC\_GETTEXT\_OPT\_CRLF（デフォルト）

DMC\_GETTEXT\_OPT\_CR

DMC\_GETTEXT\_OPT\_LF

DMC\_GETTEXT\_OPT\_U2028

DMC\_GETTEXT\_OPT\_U2029

の指定に従います。

■ シート

ワークシートの区切りには改行コードを出力します。

- 以下の形式のファイルは抽出対象外となります。
  - ① テンプレートファイル (Excel2007 除く)
  - ② アドインファイル
  - ③ ワークシートが含まれないブックファイル
  - ④ セル値レコードをなにも含まないワークシート
  - ⑤ 含まれているワークシートすべてが抽出対象で無いブックファイル
- Excel 抽出の制限事項
  - ① ヘッダとフッタでは、指定された頁番号、頁数、日付、時刻、ファイル名、シート名を抽出しません。
  - ② 「シートの保護」を設定したファイルは抽出できますが、「ブックの保護」を設定したファイルの抽出はできません。
  - ③ セルのプロパティ設定で【ユーザ定義】が選択されている場合は、表示されている文字列と抽出結果文字列が一致しない場合があります。

## 4.5 PDF

- PDF の仕様に準拠していない PDF は、動作保証ができません。
- PDF 抽出モジュールは、V4.1 より新しいモジュールを採用しております。旧バージョンと抽出結果が異なる場合があります。主な違いは以下の通り。
  - ① 抽出処理速度が従来のモジュールよりも遅くなります。
  - ② 抽出順番が従来の出力と異なる場合があります。
- PDF パッケージ (Acrobat9 : ポートフォリオ) を抽出する場合は、オプション DMC\_GETTEXT\_OPT1\_INSERTF を指定してください。指定しない場合は表紙のみの抽出となります。
- 文字と改行位置が不適当な場合があります。
- 文字間の空白が無視。または、挿入される場合があります。
- Type3 フォントは ToUnicode CMap が定義されていない場合は文字化けします。
- ユーザ定義 CMap 文字は抽出できない場合があります。または、symbol 文字は文字化けする場合があります。
- PDF 出力ソフトにより、同じ文字を重ね書きしている場合があります。この場合は、同じ文字が重複して出力されます。
- 文字のフォントが Wingdings の時、該当文字が抽出できません。
- Word、PowerPoint 文書の場合、箇条書きの行頭文字 (○●◆□■など) が、正しく抽出できない場合があります。
- Acrobat の「テキスト選択ツール」でコピーできない文字は抽出できません。

- ScanSnap で作成されたフォントが「NotDefSpecial」で、PDF 内で「Adobe-Identity-UCS」の CMap エンコーディングを参照している PDF ファイルは正しく抽出できません。
- 本ライブラリは、以下の圧縮形式に対応しております。PDF ファイル内、以下以外で圧縮されたデータは抽出できません。

FlateDecode:

LZWDecode:

ASCII85Decode

RunLengthDecode

- カスタムエンコーディングを使っている文字は、正しく抽出できません。カスタムエンコーディングを使っていると、たとえば、Adobe Reader で文字列をコピーして、エディタにペーストしたとき、文字化けします。
- PDF 内に 1 つの文字列としてまとまって入っている文字は、表示、印刷上、不連続に離れていても、そのまま 1 つの文字列として抽出されます。

## 4.6 CAD

- テキストの抽出はその格納する座標値によって順序を並べます。優先順序は Y-X-Z です。
- データ間は区切りません。(デフォルト)

## 4.7 HTML

- タグと属性を無視して、タグと属性以外の文字列を抽出します。
- <TITLE>.... </TITLE>間の文字列は、本文と区別するため、{....} のように抽出されます。
- <!-- --> で括られたコメントデータは抽出しません。

## 4.8 XML

- 文書の先頭に文字列<?xml version="1.0" .... ?>があるファイルのみ XML ファイルと認識します。(判定条件)
- XML ファイルでスタイルシートファイルを指定した場合、そのスタイルシートファイル中のテキストをテキストファイルの先頭に抽出します。
- 以下のタグ以外、タグと属性を無視して、タグと属性以外の文字列を抽出します。
  - ・ DocumentProperties
  - ・ WorksheetOptions
  - ・ PhoneticText
  - ・ ExcelWorkbook

- ・ docOleData
- ・ binData
- ・ instrText
- ・ fldData

■ 制限事項

Office2003 で作成した XML ファイル中の OLE オブジェクトは抽出しません。

- XML ファイル内で DTD が指定されていた場合、DTD が見つからない場合は DTD を無視して抽出します。

#### 4.9 Office 2001 for Macintosh, Excel98 for Macintosh

- Office 2001 for Macintosh、Excel98 for Macintosh の Apple ユーザ定義文字の抽出は保証できません。

#### 4.10 QuarkXpress

- 削除された内容が抽出される場合があります。

#### 4.11 DocuWorks

- セキュリティが設定されたファイルはエラーとなります。
- 署名されたファイルは、セキュリティ設定扱いとなり、エラーとなります。
- 太文字、影付きで修飾された文字は、文字が重複出力されます。
- テキストの抽出結果は、格納されているテキストデータの座標値によって出力順序が変わります。

見た目では、頁の先頭にある文字（段落）でも頁の先頭に出力されるとは限りません。格納されている座標値に依存します。

頁内の枠内文字列は、本文の先頭に抽出されます。

- 縦書きテキストは一文字毎に改行されます。
- DocuWorks 文書中に添付された外部ファイル<sup>\*1</sup>あるいは埋め込んだ OLE オブジェクト (MS Word/Excel/PowerPoint/RTF、PDF、一太郎、OASYS (複合型(OA3)、結合型(OA2)、オンライン形式(OAS))、AutoCAD R14/LT97 DWG、HTML、WordPerfect) のテキスト抽出ができます。(OLE 埋め込みされたファイルの抽出に対応するには、それぞれの抽出対応エンジンが必要です。)
- DocuWorks のデータを添付した場合は、Docuworks で圧縮されて格納されるため、error 3007 が返ります。
- Lotus1-2-3 は対応していません。
- 表示ページを削除した場合は、ファイルが添付されていても、テキスト抽出できません。



- \*1 対応アプリケーションのバージョンは、本仕様書の「抽出対象ファイル形式一覧」を参照してください。

#### 4.12 Visio

- 自動更新の日付、時間は正しく抽出できません。
- ヘッダ/フッタ内[p,P,t,d,D]の特殊文字は抽出しません。
- ヘッダ/フッタは以下のようにシートごと抽出します。

1 sheet 目の名称

header / footer

1 sheet 目の内容

2 sheet 目の名称

Header / footer

2 sheet 目の内容

....

- フィールド文字の抽出がされない場合があります。

#### 4.13 Outlook/Outlook Express

- 抽出には Html エンジンと、RTF エンジンを使用します。
- 添付ファイルを抽出する場合は、DMC\_GETTEXT\_OPT1\_INSERTTF をセットしてください。(オプションの詳細説明を参照)
- ストリーム出力の場合は、添付ファイルを外部フォルダに作成することができません。本文に出力します。

#### 4.14 Office 2007/Office 2010/Office 2013/Office2016/Office2019

V5 は、V4.2 に比べ、OLE 対応が大幅に拡張されました。V5.4 以降では、

Word 2007/2010/2013/2016/2019, Excel 2007/2010/2013/2016/2019, PowoerPoint 2007/2010/2013/2016/2019 に OLE で埋め込んだ Word 2003/2007/2010/2013/2016/2019, Excel 2003/2007/2010/2013/2016/2019, PowperPoint 2003/2007/2010/2013/2016/2019, Visio, PDF, 一太郎 2004 から 2019 を抽出できます。

Word 2003/2007/2010/2013/2016/2019, Excel 2003/2007/2010/2013/2016/2019, PowperPoint 2003/2007/2010/2013/2016/2019, PDF, 一太郎 2004 から 2019, DocuWorks に埋め込まれた Word 2007/2010/2013/2016/2019, Excel 2007/2010/2013/2016/2019, PowoerPoint 2007/2010/2013/2016/2019 を抽出できます。

詳しくは、「4.1.4 OLE オブジェクト抽出」にある「主要アプリケーションの OLE 対応表」を参照してください。

#### **4.15 OASYS**

- 制限事項

OLE オブジェクトとしての OASYS 文書は、DocuWorks 文書、Office 2007 文書に埋め込まれたときのみ、抽出します。

#### **4.16 OpenOffice 1.0**

- 制限事項

ファイルの判別機能のみです。テキスト抽出機能はありません。

#### **4.17 OpenOffice.org 3.1/3.2/3.3, Libre Office 3.3/3.4**

- 制限事項

OLE には対応していません。PDF に埋め込まれたときのみ抽出可能。

## 5. 付録

---

### 5.1 開発環境

参考として、以下にビルド環境を示します。

- Windows 32bit/64bit  
Microsoft Visual C++ 2010 SP1
- Solaris SPARC 32bit/64bit  
Sun C/C++ 5.8
- Solaris x86 64bit  
Sun C/C++ 5.8
- Linux 32bit/64bit  
GCC 4.1.2

### 5.2 text\_oem.h のプラットフォーム定義マクロについて

TextPorter で使う text\_oem.h には、プラットフォーム依存性を吸収するためのマクロが含まれています。

お使いのプラットフォームに合わせて、以下に示すプラットフォーム定義マクロを適切に定義することで、これらのプラットフォーム依存性を吸収するマクロが有効になります。

実際のマクロ定義の方法は、お使いのコンパイラ等に依存しますので、お使いのコンパイラ等のマニュアルを参照してください。text\_oem.h をインクルードする前に、#define で定義する方法もあります。

なお、text\_oem.h で使用しているプラットフォーム定義マクロには、現在、出荷していないプラットフォームについての記述も含まれていることをご了承ください。

プラットフォーム定義マクロの一般的な形式は、  
DMC\_<OS 名>\_<プロセッサ名>\_<ビット数>  
です。

たとえば、OS が Windows、プロセッサが Intel の x86(および AMD など互換プロセッサ)で、ビット数が 32 ビットの場合、DMC\_WINDOWS\_X86\_32 となります。

・ DMC\_WINDOWS\_X86\_32

OS	Windows
プロセッサ	Intel の x86(および AMD など互換プロセッサ)
ビット数	32 ビット

の場合に定義してください。

・ DMC\_WINDOWS\_X86\_64

OS	Windows
プロセッサ	Intel の x86(および AMD など互換プロセッサ)
ビット数	64 ビット

の場合に定義してください。

・ DMC\_LINUX\_X86\_32

OS	Linux
プロセッサ	Intel の x86(および AMD など互換プロセッサ)
ビット数	32 ビット

の場合に定義してください。

・ DMC\_LINUX\_X86\_64

OS	Linux
プロセッサ	Intel の x86(および AMD など互換プロセッサ)
ビット数	64 ビット

の場合に定義してください。

・ DMC\_SOLARIS\_SPARC\_32

OS	Solaris
プロセッサ	Sun の SPARC
ビット数	32 ビット

の場合に定義してください。

・ DMC\_SOLARIS\_SPARC\_64

OS	Solaris
プロセッサ	Sun の SPARC
ビット数	64 ビット

の場合に定義してください。

・ DMC\_SOLARIS\_X86\_64

OS	Solaris
プロセッサ	Intel の x86(および AMD など互換プロセッサ)
ビット数	64 ビット

の場合に定義してください。

・ DMC\_AIX\_POWER\_32

OS	AIX
プロセッサ	IBM の Power
ビット数	32 ビット

の場合に定義してください。

・ DMC\_AIX\_POWER\_64

OS	AIX
プロセッサ	IBM の Power
ビット数	64 ビット

の場合に定義してください。

・ DMC\_HPUX\_PARISC\_32

OS	HP-UX
プロセッサ	HP の PA-RISC
ビット数	32 ビット

の場合に定義してください。

・ DMC\_HPUX\_IA\_64

OS	HP-UX
プロセッサ	Intel Itanium など IA アーキテクチャのプロセッサ
ビット数	64 ビット

の場合に定義してください。

・ DMC\_MACOSX\_POWER\_32

OS	Mac OS X
プロセッサ	IBM の Power
ビット数	32 ビット

の場合に定義してください。

・ DMC\_MACOSX\_X86\_32

OS	Mac OS X
プロセッサ	Intel の x86(および AMD など互換プロセッサ)
ビット数	32 ビット

の場合に定義してください。

### 5.3 著作権情報

#### ■ International Components for Unicode (ICU)

This product includes software developed by International Business Machines Corporation(IBM).

URL: <http://www.icu-project.org/>

ICU License - ICU 1.8.1 and later

#### COPYRIGHT AND PERMISSION NOTICE

Copyright (c) 1995-2010 International Business Machines Corporation and others All rights reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, provided that the above copyright notice(s) and this permission notice appear in all copies of the Software and that both the above copyright notice(s) and this permission notice appear in supporting documentation.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE BE LIABLE FOR ANY CLAIM, OR ANY SPECIAL INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

#### ■ zlib

This product includes software developed by Jean-loup Gailly and Mark Adler.

URL: <http://zlib.net>

/\* zlib.h -- interface of the 'zlib' general purpose compression library  
version 1.2.5, April 19th, 2010

Copyright (C) 1995-2010 Jean-loup Gailly and Mark Adler

This software is provided 'as-is', without any express or implied warranty. In no event will the authors be held liable for any damages arising from the use of this software.

Permission is granted to anyone to use this software for any purpose, including commercial applications, and to alter it and redistribute it freely, subject to the following restrictions:

1. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.
2. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.
3. This notice may not be removed or altered from any source distribution.

Jean-loup Gailly  
Mark Adler

\*/

## ■ unrar

This product is not RAR compatible archiver. This product includes software developed by Eugene Roshal.

URL: <http://www.rarlab.com/>

The source code of unRAR utility is freeware. This means:

1. All copyrights to RAR and the utility unRAR are exclusively owned by the author - Eugene Roshal.
2. The unRAR sources may be used in any software to handle RAR

archives without limitations free of charge, but cannot be used to re-create the RAR compression algorithm, which is proprietary. Distribution of modified unRAR sources in separate form or as a part of other software is permitted, provided that it is clearly stated in the documentation and source comments that the code may not be used to develop a RAR (WinRAR) compatible archiver.

3. The unRAR utility may be freely distributed. No person or company may charge a fee for the distribution of unRAR without written permission from the copyright holder.
4. THE RAR ARCHIVER AND THE UNRAR UTILITY ARE DISTRIBUTED "AS IS".

NO WARRANTY OF ANY KIND IS EXPRESSED OR IMPLIED. YOU USE AT YOUR OWN RISK. THE AUTHOR WILL NOT BE LIABLE FOR DATA LOSS, DAMAGES, LOSS OF PROFITS OR ANY OTHER KIND OF LOSS WHILE USING

OR MISUSING THIS SOFTWARE.

5. Installing and using the unRAR utility signifies acceptance of these terms and conditions of the license.
6. If you don't agree with terms of the license you must remove unRAR files from your storage devices and cease to use the utility.

Thank you for your interest in RAR and unRAR.

Eugene Roshal

## ■ LHa

This product includes software developed by LHa.

URL: <http://lha.sourceforge.jp/>

.

以下の条件で、再配布、転載、改変を許可します。

1. 著作権表示を削除しないこと。
2. 配布内容については、
  - a. 配布の際に存在する内容(すなわちソースコード、ドキュメント、プログラマーへの手引きなど)が再配布されたものの中に必ず存在すること。改変されているならば、それを明示したドキュメントを用意すること。
  - b. LHa に対する付加価値が付けられて再配布される場合に



はそれらもできるだけ含めるよう努力すること。また、その際には付加価値が付けられていることを明示したドキュメントを用意すること。

- c. バイナリのみ配布は許されない。(付加価値のものも含む)
- 3. 最新版の配布に務めること。(義務はない)
- 注. なお、ネットでの配付は自由であるが、ネットにアクセスできない方（雑誌および、CD-ROM などによる）配付は、配付前にこちらに E-Mail をお願いします。配付前にできない際には、後日必ず E-Mail をお願いします。
- 4. このプログラムの存在や使用したことによって生じた損害は全く保証しない。
- 5. 作者は、このプログラムに不備があっても、それを訂正する義務を負わない。
- 6. このプログラムの一部、または全部を他のプログラムに組み込んで利用してもかまわない。この場合、そのプログラムは LHa ではなく、LHa と名乗ってはいけない。
- 7. 商利用に関しては、上記の条件に加え、下記の条件のもとにこれを認める。
  - a. このプログラムをメインとする商利用は禁止する。
  - b. 商利用の相手がこのプログラムの使用者として不適切と判断した場合には配布しない。
  - c. インストールの手段として使用する場合、このプログラムを使うことを相手に強制しない。この場合、商利用者が作業を行う。また、そのときの損害は、商利用者が全責任を負う。
  - d. 商利用を付加価値として行いこのプログラムを使用する場合、商利用者は、そのサポートを行う。

## 5.4 サンプルアプリケーションの使用法

C/C++で書いたライブラリ利用のサンプルが同梱されています。

名前は、app\_が先頭についています(Windows なら、app\_ww など)。

製品パッケージには、ソースコードも入っているので、参考にしてください。

サンプルは、無保証、無サポートです。また、予告なく、以前のバージョンとは非互換な修正が行われる可能性もあります。あくまで、ライブラリ利用のサンプルである点、ご承知おきください。

### 実行する前に

1. Windows 系は、\*.dll、\*.dat、app\_ww (to\_com\_vb) を同じフォルダに保存してください。格納したフォルダを、環境変数 PATH に指定しておいてください。
2. Unix 系の場合は、\*.so\*、\*.dat、app\_\*を同じディレクトリに保存してください。格納したディレクトリを、環境変数 LD\_LIBRARY\_PATH など、そのプラットフォームが共有ライブラリ検索に使用する環境変数に指定しておいてください。
3. 文字コード変換テーブルの保存先を、環境変数 DMC\_TBLPATH で指定することができます。指定しないと、ライブラリを格納したディレクトリにある base2 を想定します。

(以下は app\_ww のみ示します)

**app\_ww を実行すると、ヘルプとして**

*Usage: app\_ww input-app-file [-o output-txt-file] [-t target-dir] [-g group-name] [-e big-endian] [-d default-language-name] [-p option] [-m do-multi-thread] [-n pages] [-f function-number] [-w password] [-s size] [-c csv\_c] [-hi thread-interval] [-hn thread-number] [-a param-file]*

(以下、省略)

を表示します。

### 1. 入力アプリケーションファイルの指定

書式 : app\_ww input-app-file

【例】

- |                |   |                      |
|----------------|---|----------------------|
| 一つのファイルを抽出する場合 | → | app_ww t2.doc        |
| 複数のファイルを抽出する場合 |   |                      |
| ・異なるフォーマット     | → | app_ww t2.xls t3.doc |
| ・ワイルドカード       | → | app_ww *.doc         |

- ・ 指定フォルダ／ディレクトリ内のすべて → `app_ww ¥textporter¥test¥*`

## 2. 抽出先ファイル名の指定

書式 : `app_ww input-app-file [-o output-txt-file]`

`output-txt-file` の指定がない場合は、自動的に元のファイル名+.txt がtxtfile 名になります。

## 3. 抽出先ファイルの格納場所の指定

書式 : `app_ww input-app-file [-t target-dir]`

`target-dir` の指定がない場合は、`app_**`と同じディレクトリに格納されます。

## 4. オプション

**default-language-name:** (抽出元ファイルの言語指定)

```

en
jp      (default)
cn
tw
ko

```

オプション 指定	対応言語
en	English
jp	Japanese
cn	Simplified Chinese
tw	Traditional Chinese
ko	Korean

**group-name:** (抽出先文字集合名指定)

```

EUC-JP
EUC-JP-FIX
ISO-10646-UCS-2
ISO-10646-UCS-4
ISO-2022-JP
ISO-8859-1
Shift_JIS      (default)
UTF-16
UTF-8
WINDOWS31J
ChineseGBK(V4 must use GBK)

```

ChineseBIG5(V4 must use Big5)  
 GB18030  
 KoreanKSC(V4 must use KS\_C\_5601-1987)  
 Shift\_JIS-2004  
 ISO-2022-JP-2004  
 EUC-JIS-2004

**big-endian:** (エンディアン指定)

0: little endian  
 1: big endian (default)

**option:** (オプション指定、文字列で指定してください)

オプションの意味は、「3.4.2 テキスト抽出関数」を参照してください。

DMC\_GETTEXT\_OPT\_KEISEN  
 DMC\_GETTEXT\_OPT\_TAG  
 DMC\_GETTEXT\_OPT\_CRLF (default)  
 DMC\_GETTEXT\_OPT\_CR  
 DMC\_GETTEXT\_OPT\_LF  
 DMC\_GETTEXT\_OPT\_U2028  
 DMC\_GETTEXT\_OPT\_U2029  
 DMC\_GETTEXT\_OPT\_SHEET  
 DMC\_GETTEXT\_OPT\_RUBI  
 DMC\_GETTEXT\_OPT\_PWD  
 DMC\_GETTEXT\_OPT\_OLE  
 DMC\_GETTEXT\_OPT\_OLE1  
 DMC\_GETTEXT\_OPT\_OLE2  
 DMC\_GETTEXT\_OPT\_OLE3  
 DMC\_GETTEXT\_OPT\_OUT  
 DMC\_GETTEXT\_OPT\_LOOP  
 DMC\_GETTEXT\_OPT\_SHEET1  
 DMC\_GETTEXT\_OPT\_CELL  
 DMC\_GETTEXT\_OPT\_SIZE  
 DMC\_GETTEXT\_OPT\_CSV1  
 DMC\_GETTEXT\_OPT\_CSV2  
 DMC\_GETTEXT\_OPT\_PDFSYM  
 DMC\_GETTEXT\_OPT\_OWNERPWD1

DMC\_GETTEXT\_OPT\_OWNERPWD2  
 DMC\_GETTEXT\_OPT\_OWNERPWD3  
 DMC\_GETTEXT\_OPT\_OWNERPWD4  
 DMC\_GETTEXT\_OPT1\_OWNERPWD5  
 DMC\_GETTEXT\_OPT\_ENDCODE  
 DMC\_GETTEXT\_OPT\_NULL  
 DMC\_GETTEXT\_OPT\_SHFTAG  
 DMC\_GETTEXT\_OPT\_SHFHEAD  
 DMC\_GETTEXT\_OPT1\_TEMP  
 DMC\_GETTEXT\_OPT1\_INSERTF  
 DMC\_GETTEXT\_OPT1\_INSERTF1  
 DMC\_GETTEXT\_OPT1\_INSERTF2  
 DMC\_GETTEXT\_OPT1\_INSERTF3  
 DMC\_GETTEXT\_OPT1\_COMPRESS  
 DMC\_GETTEXT\_OPT1\_COMPRESS1  
 DMC\_GETTEXT\_OPT1\_COMPRESS2  
 DMC\_GETTEXT\_OPT1\_COMPRESS3  
 DMC\_GETTEXT\_OPT1\_COMPRESS4  
 DMC\_GETTEXT\_OPT1\_TRACK  
 DMC\_GETTEXT\_OPT1\_COMPRESS5  
 DMC\_GETTEXT\_OPT1\_INSERTF4  
 DMC\_GETTEXT\_OPT1\_TXCONV  
 DMC\_GETTEXT\_OPT1\_TXCONV2  
 DMC\_GETTEXT\_OPT1\_OUTPUT\_RAW\_NL  
 DMC\_GETTEXT\_OPT1\_QUOTE\_QQ

**do-multi-thread:** (シングルスレッドのテスト／マルチスレッドのテスト指定)

0: single thread test      (default)  
 1: multi-thread test

**pages:** (0は文書の頁数を取得する。1,2,3...は抽出したい頁番号の指定)

0: get page count      (default)  
 1,2,3.....: page number

【注意】 DMC\_GetPageText\_V5、DMC\_GetPwdPageText\_V5 を利用する場合有効

**function-number:** (テキスト抽出機能関数の指定、3~5、11~13 は PDF のみ有効)

- 0: test DMC\_GetText\_V5 (default)
- 1: test DMC\_GetPageText\_V5
- 2: test DMC\_GetProperty\_V5
- 3: test DMC\_GetPwdText\_V5
- 4: test DMC\_GetPwdPageText\_V5
- 5: test DMC\_GetPwdProperty\_V5
- 6: test DMC\_GetTextStream\_V5
- 7: test DMC\_GetPageTextStream\_V5
- 8: test DMC\_GetText\_V4
- 9: test DMC\_GetPageText\_V4
- 10: test DMC\_GetProperty\_V4
- 11: test DMC\_GetPwdText\_V4
- 12: test DMC\_GetPwdPageText\_V4
- 13: test DMC\_GetPwdProperty\_V4
- 14: test DMC\_GetTextStream\_V4
- 15: test DMC\_GetPageTextStream\_V4

**password:** (パスワードを入力してください)

【注意】DMC\_GetPwdText\_V5、DMC\_GetPwdPageText\_V5、DMC\_GetPwdProperty\_V5  
を利用する場合有効

**size:** (テキスト取りだし最大 Byte 数を入力してください)

【注意】オプション DMC\_GETTEXT\_OPT\_SIZE を指定した場合有効

**csv\_c:** (Excel、Lotus1-2-3 の列間区切り文字コードを 10 進数で入力してください)

指定可能な文字：1 Byte の ASCII 文字（制御文字を除く）のみ

【注意】オプション DMC\_GETTEXT\_OPT\_CSV2 を指定した場合有効

**thread-interval:** (スレッドの実行間隔を秒単位の 10 進数で入力してください)

デフォルトは、3 秒です。

オプション -m 1 で、マルチスレッドテストを指定した場合有効

**thread-number:** (スレッド数を 10 進数で入力してください)

デフォルトは、50 です。

オプション -m 1 で、マルチスレッドテストを指定した場合有効

**param-file:** (パラメータファイル指定)

コマンドラインのパラメータをファイルに記述できます。

それをパラメータファイルと呼びます。

パラメータファイルを、-a で指定することができます。

パラメータファイルの形式は以下の通りです。

---

```
#で始まる行はコメントです。
各パラメータは改行で区切ります。
例えば
app_ww C:\¥in¥sample.zip -p DMC_GETTEXT_OPT1_COMPRESS
-p DMC_GETTEXT_OPT1_COMPRESS2 -t C:\¥out
同じ機能は
param.dmc
#-----
C:\¥in¥sample.zip
-p DMC_GETTEXT_OPT1_COMPRESS
-p DMC_GETTEXT_OPT1_COMPRESS2
-t C:\¥out
#-----
として
app_ww -a param.dmc
で実現できます。
パラメータファイルは、ANSI、UTF-16、UTF-8 で書くことが可能です。
UTF-16、UTF-8 のときは、必ず BOM(Byte Order Mark)を入れてください。
Windows 版では、パラメータファイルのエンコーディングでファイル名が表現されている
ものとして動きます。これにより、ANSI では表現できないファイル名を扱うことができます。
-e と -m では、yes/no、on/off など使えます。
-f では、DMC_GetText_V5 など、関数の名前も使えます。
```

---

## 5.5 To\_com\_vcs の使用法

C#で書いたライブラリ利用のサンプルが同梱されています。

製品パッケージには、ソースコードも入っているので、参考にしてください。

実行するには、あらかじめ regsvr32 で To\_com.dll を登録しておく必要があります。

パラメータやオプションについては、「5.4 サンプルアプリケーションの使用法」も参考にしてください。

サンプルは、無保証、無サポートです。また、予告なく、以前のバージョンとは非互換な修正が行われる可能性もあります。あくまで、ライブラリ利用のサンプルである点、ご承知おきください。

### 起動画面

Input application file	抽出元ファイルを指定します。
Output text file	出力先テキストファイルを指定します。
Function	テストする API を指定します。
Pages	ページ抽出する場合のページ番号を指定します。
Password	パスワード付 PDF ファイルを抽出する場合の解除パスワードを指定します。
Character encoding	出力の文字エンコーディングを指定します。
Default language	入力のデフォルト言語を指定します。



Big endian

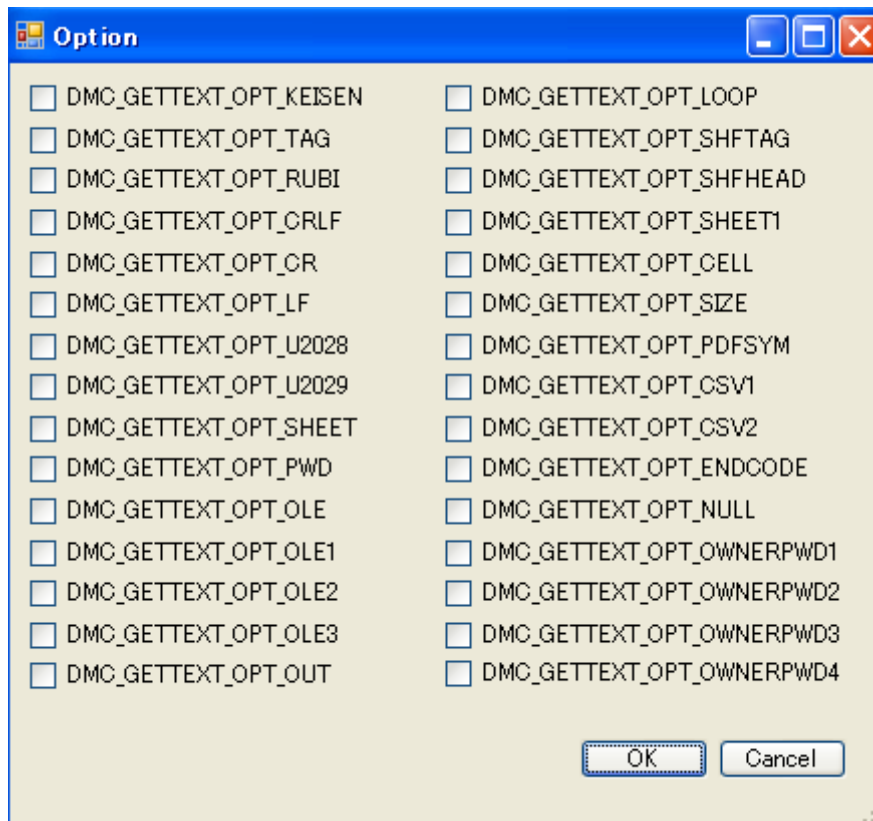
Big endian かどうかを指定します。

Option

オプションを指定します。

「Setup」 ボタンを押すと、一覧から選べます。

### Option の設定画面

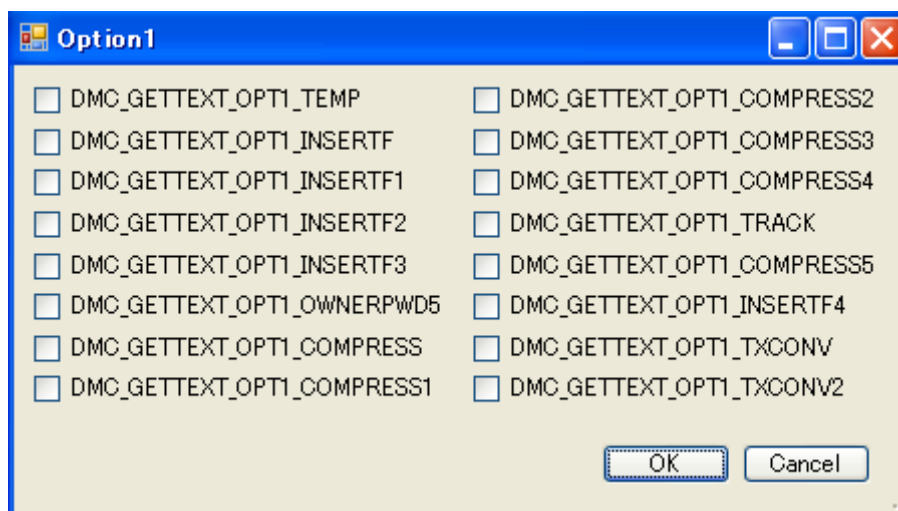


Option1

オプション 1 を指定します。

「Setup」 ボタンを押すと、一覧から選べます。

### Option1 の設定画面



Size	出力サイズを指定します。
CSV delimita char code	CSV 出力の区切り文字を指定します。
Convert	テキスト抽出を実行します。
End	To_com_vcs を終了します。

## 5.6 liview の使用法

liview は、ライセンス管理ファイル dmc\_txli.dat をテキスト化するツールです。

### 実行する前に

実行前の準備は、「5.4 サンプルアプリケーションの使用法」の「実行する前に」を参照してください。

### liview -?を実行すると、ヘルプとして

*Usage: liview [filename]*

*filename - full path name of the license file.*

*if the filename is not specified, liview will use dmc\_conf to search file dmc\_txli.dat*

を表示します。

**filename:**ライセンス管理ファイル名を指定してください。

filename が指定されない場合、dmc\_conf で指定したライセンス管理ファイルの格納パスにある dmc\_txli.dat を検索します。

### 表示形式

liview を実行すると、ユーザ情報、対応するプラットフォーム、契約情報、対応フォーマット、言語などを下記のように表示します。

評価版の場合、LicenseType は Eval になります。

Expire の日付は、保守期間の終了日を示します。テキスト抽出ができなくなる有効期限ではありません。保守期間終了日以降にリリースされたものには、バージョンアップできません。

『例』

```
TextPorter License Information
Company: アンテナハウス (株)、テスト用
Section:
UserName:
Version: 5.3
Platform: Windows 32bit
LicenseType: Stand alone
Expire: No limit
```

0 HTML

Japanese support: Yes  
English support: Yes  
Simplified Chinese support: Yes  
Traditional Chinese support: Yes  
Korean support: Yes

1 PDF1.2

Japanese support: Yes  
English support: Yes  
Simplified Chinese support: Yes  
Traditional Chinese support: Yes  
Korean support: Yes

...

25 Excel 2007

Japanese support: Yes  
English support: Yes  
Simplified Chinese support: Yes  
Traditional Chinese support: Yes  
Korean support: Yes

## 5.7 パスワード付き PDF 文書のテキスト抽出

本ライブラリでは、パスワード付き PDF 文書からテキストを抽出することができます。

PDF のパスワードには、セキュリティ設定を変更するパスワード（オーナーパスワード、Owner Password）、文書を開くパスワード（ユーザパスワード、User Password）の 2 種類のパスワードがあります。

パスワードの種類によって、呼び出す関数やオプションが違います。

### ★注意事項

PDF では、セキュリティに関する処理は、セキュリティハンドラがつかさどります。セキュリティハンドラは自由に定義できますが、本ライブラリで対応しているのは、あらかじめ定義されている標準セキュリティハンドラのみです。独自のセキュリティハンドラを定義している場合は対応できません。

標準のセキュリティハンドラについての詳しい解説は、

[http://www.adobe.com/devnet/pdf/pdf\\_reference\\_archive.html](http://www.adobe.com/devnet/pdf/pdf_reference_archive.html)

Adobe PDF Reference Archives

にある「PDF Reference, Sixth Edition, version 1.7」の「3.5.2 Standard Security Handler」あるいは、

[http://www.adobe.com/devnet/pdf/pdf\\_reference.html](http://www.adobe.com/devnet/pdf/pdf_reference.html)

PDF Reference and Adobe Extensions to the PDF Specification

にある「Document Management – Portable Document Format – Part 1: PDF 1.7, First Edition」の「7.6.3 Standard Security Handler」を参照してください。

### オーナーパスワードが設定されている PDF 文書

- オーナーパスワードが設定された文書は、通常の文書と同じように DMC\_GetText\_V5 関数などを使って、テキストを抽出できます。

テキスト抽出オプションの指定により、テキスト抽出の制御が可能です。

詳しくは、付録「セキュリティ設定した PDF のテキスト抽出制御仕様」を参照してください。

### ユーザパスワードが設定されている PDF 文書

- ユーザパスワードが設定された文書の抽出時、DMC\_GetText\_V5, DMC\_GetTextStream\_V5, DMC\_GetPageText\_V5, DMC\_GetPageTextStream\_V5, DMC\_GetProperty\_V5 関数では、エラーコード 5000(ユーザパスワード付き PDF ファイル)を返します。

- ユーザパスワード付き文書进行处理する場合は、ユーザインターフェースからパスワードを指定して、DMC\_GetPwdText\_V5 関数など、パスワード付きファイルの関数と、オプション DMC\_GETTEXT\_OPT\_PWD を使ってください。

## 5.8 セキュリティ設定した PDF のテキスト抽出制御仕様

PDF では、オーナーパスワードによって、セキュリティ設定(パーミッション設定)を行うことができます。

本ライブラリは、オーナーパスワードによってセキュリティ設定(パーミッション設定)をされた PDF ファイルについては、何もオプションを指定しなくても、通常の文書と同じように DMC\_GetText\_V5 関数などを使って、テキストを抽出できます。

しかし、パーミッションの種類によっては、テキストを抽出したくない場合もあります。たとえば、変更が禁止されている PDF ファイルについては、テキストを抽出しないといった場合です。

その場合、ここで述べるオプションを使って、テキストを抽出せず、エラーを意図的に起こすことができます。

以下で説明するパーミッションのビットの意味について、詳しい解説は、

[http://www.adobe.com/devnet/pdf/pdf\\_reference\\_archive.html](http://www.adobe.com/devnet/pdf/pdf_reference_archive.html)

Adobe PDF Reference Archives

にある「PDF Reference, Sixth Edition, version 1.7」の「TABLE 3.20 User access permissions」

あるいは、

[http://www.adobe.com/devnet/pdf/pdf\\_reference.html](http://www.adobe.com/devnet/pdf/pdf_reference.html)

PDF Reference and Adobe Extensions to the PDF Specification

にある「Document Management – Portable Document Format – Part 1: PDF 1.7, First Edition」の「Table 22 – User access permissions」

を参照してください。

PDF に設定されるパーミッションの情報は、ビット毎に意味が決まっています。

大きく次に分類できます。

- ・ 印刷を許可する(bit 3)
- ・ 変更を許可する(bit 4)
- ・ 内容のコピーを許可する(bit 5)

です。

これらは、さらに細分化されています(現在の本ライブラリのオプションは、これらのビットすべてに対応しているわけではありません)。

印刷を許可するビットには、bit 3 のほかに、次があります。

- ・ 高解像度印刷を許可する(bit 12)

変更の許可するビットには、bit 4 のほかに、次があります。

- ・ 注釈の追加、変更、フォームフィールドの入力を許可する(bit 6)
- ・ bit 6 が 0 でも、既存のフォームフィールドの入力を許可する(bit 9)

- ・ ページの挿入、削除、回転を許可する(bit 11)

内容のコピーを許可するビットには、bit 5 のほかに、次があります。

- ・ アクセシビリティのためにコピーを許可する(bit 10)

ここで説明するオプションは、これらのビットがクリアされているとき(0 のとき)、すなわち許可がないときに、エラーを発生させるオプションです。

ビットとオプションの対応は、次のとおりです(bit 9, bit 11, bit 12に対応するオプションはありません,.)。

ビット	対応するオプション
bit 3	DMC_GETTEXT_OPT_OWNERPWD1
bit 4	DMC_GETTEXT_OPT_OWNERPWD2
bit 5	DMC_GETTEXT_OPT_OWNERPWD3
bit 6	DMC_GETTEXT_OPT_OWNERPWD4
bit 10	DMC_GETTEXT_OPT1_OWNERPWD5

たとえば、印刷を許可する bit 3 が 0 のときに、すなわち、印刷が許可されていないときに、DMC\_GETTEXT\_OPT\_OWNERPWD1 を指定すると、エラーコード 5003(Owner Password によるセキュリティ設定(パーミッション設定)がされたファイル)を発生させることができます。



以下の表は、セキュリティの設定と、エラーを起こすためのオプションを、暗号化レベルや Acrobat の種類によって分類して、まとめたものです。

対応するオプションが「なし」になっているところは、前述のように、bit 9, bit 11, bit 12 に対応するオプションがない部分です。

**暗号化レベル: 40-bit RC4 (Acrobat 3.x, 4.x)**

セキュリティ設定	設定内容	対応するオプション
印刷	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
文書の変更	許可しない	DMC_GETTEXT_OPT_OWNERPWD2
テキストとグラフィックの選択	許可しない	DMC_GETTEXT_OPT_OWNERPWD3
注釈とフォームフィールドの追加と変更	許可しない	DMC_GETTEXT_OPT_OWNERPWD4

**暗号化レベル: 40-bit RC4 (Acrobat 6、Acrobat3.0 以降)**

セキュリティ設定	設定内容	対応するオプション
印刷を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
	高解像度	なし
変更を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	フォームフィールドの入力と署名	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	注釈の作成、フォームフィールドの入力と署名	DMC_GETTEXT_OPT_OWNERPWD2
	ページの抽出を除くすべての操作	なし
テキスト、画像、およびその他の内容のコピーとアクセシビリティを有効にする	チェックしない	DMC_GETTEXT_OPT_OWNERPWD3

## 暗号化レベル：128-bit RC4 (Acrobat 5)

セキュリティ設定	設定内容	対応するオプション
印刷	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
	低解像度	なし
	すべて許可	なし
変更を許可	なし	DMC_GETTEXT_OPT_OWNERPWD2
	文書アセンブリのみ	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	フォームフィールドの入力または署名のみ	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	注釈の作成、フォームフィールドの入力または署名	DMC_GETTEXT_OPT_OWNERPWD2
	編集、注釈及びフォームフィールドの作成	なし
内容のコピーと抽出を許可	チェックしない	DMC_GETTEXT_OPT_OWNERPWD3
アクセシビリティを有効にする	チェックしない	DMC_GETTEXT_OPT1_OWNERPWD5

## 暗号化レベル：128-bit RC4 (Acrobat 6、Acrobat5 以降)

セキュリティ設定	設定内容	対応するオプション
印刷を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
	低解像度 (150dpi)	なし
	高解像度	なし
変更を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	ページの挿入、削除、回転	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	フォームフィールドの入力と署名	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	注釈の作成、フォームフィールドの入力と署名	DMC_GETTEXT_OPT_OWNERPWD2
	ページの抽出を除くすべての操作	なし
テキスト、画像、およびその他の内容のコピーを有効にする	チェックしない	DMC_GETTEXT_OPT_OWNERPWD3
スクリーンリーダーデバイスのテキストアクセスを有効にする	チェックしない	DMC_GETTEXT_OPT1_OWNERPWD5

## 暗号化レベル: 128-bit RC4 (Acrobat 6、Acrobat6 以降)

セキュリティ設定	設定内容	対応するオプション
印刷を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
	低解像度 (150dpi)	なし
	高解像度	なし
変更を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	ページの挿入、削除、回転	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	フォームフィールドの入力と署名	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	注釈の作成、フォームフィールドの入力と署名	DMC_GETTEXT_OPT_OWNERPWD2
	ページの抽出を除くすべての操作	なし
テキスト、画像、およびその他の内容のコピーを有効にする	チェックしない	DMC_GETTEXT_OPT_OWNERPWD3
スクリーンリーダーデバイスのテキストアクセスを有効にする	チェックしない	DMC_GETTEXT_OPT1_OWNERPWD5
書式なしテキストのメタデータを有効にする	チェックする／しない	なし

## 暗号化レベル: 128-bit AES (Acrobat 7、Acrobat7 以降)

セキュリティ設定	設定内容	対応するオプション
印刷を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
	低解像度 (150dpi)	なし
	高解像度	なし
変更を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	ページの挿入、削除、回転	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	フォームフィールドの入力と既存の署名フィールドに署名	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	注釈の作成、フォームフィールドの入力と既存の署名フィールドに署名	DMC_GETTEXT_OPT_OWNERPWD2
	ページの抽出を除くすべての操作	なし
テキスト、画像、およびその他の内容のコピーを有効にする	チェックしない	DMC_GETTEXT_OPT_OWNERPWD3
スクリーンリーダーデバイスのテキストアクセスを有効にする	チェックしない	DMC_GETTEXT_OPT1_OWNERPWD5

## 暗号化レベル: 256-bit AES (Acrobat 9、Acrobat9 以降)

セキュリティ設定	設定内容	対応するオプション
印刷を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD1
	低解像度 (150dpi)	なし
	高解像度	なし
変更を許可	許可しない	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	ページの挿入、削除、回転	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	フォームフィールドの入力と既存の署名フィールドに署名	DMC_GETTEXT_OPT_OWNERPWD2 または DMC_GETTEXT_OPT_OWNERPWD4
	注釈の作成、フォームフィールドの入力と既存の署名フィールドに署名	DMC_GETTEXT_OPT_OWNERPWD2
	ページの抽出を除くすべての操作	なし
テキスト、画像、およびその他の内容のコピーを有効にする	チェックしない	DMC_GETTEXT_OPT_OWNERPWD3
スクリーンリーダーデバイスのテキストアクセスを有効にする	チェックしない	DMC_GETTEXT_OPT1_OWNERPWD5

## 5.9 パスワード付き Microsoft Office, 一太郎のテキスト抽出

パスワード付きの Microsoft Office, 一太郎ファイルについて、ファイル識別、テキスト抽出を行った場合の結果を、参考情報として、

「text-extraction-of-password-protectd-msoffice-ichitaro.xlsx」

に、まとめています。

このファイルは、本マニュアルと同じフォルダにあります。

この結果は、本ライブラリの現在の動作を示しており、今後の製品改良に伴い、予告なく変わる可能性があります。