メールと圧縮ファイルの階層展開ライブラリ

デ変研MFXライブラリ

MFXLibrary Dehenken

Dehenken

メールの変換道

Dehenkenでは、各種フォーマットの文書ファイルから安全にテキストデータを取り出す技術を大切にしています。さまざまな形式のフォーマットから、精度良く・安全に取り出す継続的品質・性能向上活動を「変換道」と呼んでいます。

●多様なメールファイルからのメール情報の取り出し

本ライブラリは、さまざまなメーラで作成された1つのEMLファイルについて、ヘッダとボディとその添付ファイルを取り出します。添付ファイルはさらに添付展開し、そこに圧縮ファイルがあればそれを展開して、最終的に1つのファイルになるまで展開します(展開レベル指定による制限可)。対応している添付ファイルのエンコードは、uuencode(MIME形式／本文埋込方式)、base64、quoted-printable、binhexです。

また、UNIXのmailコマンドでいうmboxファイルのような複合メール形式について、1つずつのメールを取り出し展開するAPIを用意しています。対応している復号メール形式は、mbox、thunderbird、Outlook、Outlook Express、Becky! 等のソフトウェアで作成されたものです(Becky!のみ添付ファイルの展開は未対応)。

●さまざまな圧縮ファイルからのファイルの取り出し

本ライブラリは、さまざまな圧縮ソフトで作成された圧縮ファイルを展開します。展開は1つのファイルから再帰的に行い、最終的には1つのファイルのレベルまで掘り下げます。圧縮したアーカイブファイルは、zip(winzip／pkzip)、lha(自己解凍形式を含む)、gzip、tar+gz、rar、bzip2です。圧縮していないアーカイブファイルはtar / gnutarです。

※各種パスワード付は除きます。

●メールヘッダを変換して提供

メールのヘッダ部分を変換して取り出しをすることができます。メールヘッダのCc: などに多くの配信先を指定した場合には、複数行におよぶ場合があります。本ライブラリは、複数行に及んだ行は連結して1行にして結果の提供を行います。また、ヘッダにB-エンコード、あるいはQ-エンコードがあれば、デコードして結果の提供を行います。また、メールの取り出し時の文字コードの指定(後述のlanguageによる)ができますので、ユーザが得られるヘッダ情報は指定された漢字コードで得ることができます。

●メール本文を指定の文字コードに変換

メール本文のテキスト情報を指定の文字コードで取り出すことができます。メール本文が途中で破損し、バイナリデータが混入していた場合、破損部以外の文字情報を可能な限り取り出します。また、メールの途中の文字コードに異常データ(0x00-0x1fで改行とタブを除く)が含まれていてもそれを取り除く処理をします。

●メール添付情報と、付属するMIME属性を得られます

メールに添付されている複数の添付ファイルの1つ1つを順に取り出すAPIを用意しています。また、このときのメールの添付に付属する添付ファイル名の情報を取り出すことができます。

●ユニコードの対応

本ライブラリは、日本語文字コードの2つの方式Windows-31J(CP932)とJIS X0213:2004に対応しています。これらの文字コード形式を指示した上で、JIS (ISO-2022-JP) / EUC-JP / Shift_JIS / UTF-8 / UTF-16の変換を行います。

本ライブラリは、content-typeの指定によって、テキストの文字コードの判別を優先的に行うようにしています。デ変研MFXライブラリ multi-language版では、ヨーロッパ圏の文字コードISO-8859や、アジアの中国語、韓国語についても対応しています。(GB2312 / GB18030 / KSC 5601 / Big5 等)。

●開発生産性を支援する情報の提供と処理の工夫

本ライブラリを使用して開発する技術者のために、開発生産性を高めるために様々な情報の提供と工夫をしています。

- ●サンプルソースを用意しており開発時の参考にしていただけます。
- ●ファイルをAPIに与えるだけで、ライブラリが自動判別して動作します。
- ●ファイルの判定にはサフィックス(接尾文字)の情報を使いません。
- ●ファイル名指定方式とファイル内容メモリ渡しの2つのインターフェースをご用意しています。

一般に、プログラム経験のある方であれば、容易に使い方をご理解いただけます。

●コールバック関数の指定

ユーザは、ライブラリにコールバック関数を登録することができます。本ライブラリは、大量の圧縮ファイル展開のために長時間処理を占有する場合がありますが、コールバック関数を用いて処理の途中に、中断やWaitの処理を入れることができ、CPUの占有を防ぐことができます。

●マルチスレッドに対応

本ライブラリは、CPU数やコア数に応じてテキスト抽出の分散処理による速度の向上ができるよう、マルチスレッドに対応しています(マルチスレッドセーフ)。

●破損ファイル対象時の安全性(Broken Files Safeness)

文書ファイルやテキストファイルは、保存時や複製時にデータの破損事故が発生する場合があります。本ソフトウェアが破損したファイルであっても、安全に抽出を継続したり、中断したりできることを確認するために、意図的に破壊したファイルで網羅的に動作確認を実施しています。当社のこのような破損ファイルによる安全確認をBroken Files Safenessと呼んでいます。

Dehenken

株式会社 データ変換研究所 Dehenken Limited

本　　社 〒604-8155 京都市中京区錦小路通室町東入占出山町 308 ヤマチュウビル 1F
TEL 075-254-8780 FAX 075-254-8790

横浜営業所 〒231-0048 神奈川県横浜市中区蓬莱町 2-4-7 澤田聖徳ビル 204

URL : http://www.dehenken.co.jp/ E-Mail : info_ml@dehenken.co.jp

品質マネジメントシステム ISO 9001:2008 の認証取得

株式会社データ変換研究所は、2011年9月27日付けで全社統一の品質マネジメントシステムとしてDNV GL より ISO-9001:2008 の認証を取得しました。(現在は2015年版に移行)

認証の対象は「ソフトウェアプロダクトのデザイン・開発・製造」です。

QUALITY SYSTEM CERTIFICATION DNV·GL ISO 9001

Certificate No. : 02523-2011-AQ-KOB-RvA
Initial certification date : 27 September, 2011
Valid : 27 September, 2017 - 27 September, 2020

開発生産性を高める、メールの使用環境を提供

デ変研MFXライブラリ(以下、本ライブラリ)を用いてアプリケーションを開発するOEMユーザを支援するために、開発しやすく工夫した様々な提供を行います。API使用時のサンプルソース(再利用できる使用権付与)もご提供しています。
展開した１つのメールファイルのファイル名を指定してAPIに与え、ライブラリがメール形式や圧縮形式を自動判別し展開します。このときのファイルの判定には拡張子(サフィックスまたは接尾文字)の情報を使いません。

本ライブラリで対応しているE-Mail形式の対応仕様について

●１つのメール展開

EML / OutlookのMSG形式
(EMLは、E-Mail形式のことで、RFC822に準拠したものをいいます)

●複合メール形式

複合メール形式	ヘッダ展開	本文展開	添付展開
Mbox形式 ※1※2	○	○	○
PST (Outlook 32 / 64 bit) 形式※1※3	○	○	○
DBX (Outlook Express) 形式※4	○	○	○
Becky!形式	○	○	×

※1…複合メール形式から展開した１通のE-Mail形式のサイズは2Gbyteまで、
またメールの通数は2147483647(signed intの最大値)までとなります。
※2…MBOX形式に対応しているアプリの１つの例として、Thunderbirdがあります。
※3…Outlook2002は、2Gbyte以上のサイズのファイルへは対応していません。
※4…2Gbyte以上のサイズのファイルへは対応していません。

●圧縮ファイル

zip (winzip / pkzip : 圧縮形式 / 自己解凍形式)
lha (lh1 / lh5 / lh6 / lh7 : 圧縮形式 / 自己解凍形式)
tar+gzip / tgz / gzip
rar (圧縮形式 / 自己解凍形式)
bzip2
7z (圧縮形式 / 自己解凍形式)
※それぞれの圧縮形式において、パスワード付きのものを除きます。

●アーカイブ形式

tar / gnutar

●メールヘッダのエンコード

メールヘッダのエンコードは、MIME Q/B 及び RFC2231に対応しています。

●メール本文のエンコード

メール本文のエンコードは、エンコードなし(text/plain)
base64、quoted-printableに対応しています。

●添付ファイルのエンコード

添付ファイルのエンコードは、uuencode(MIME形式/本文埋め込み形式)、base64、quoted-printable、binhexに対応しています。

●パート判別

E-Mail形式におけるパートの判別方式は、FromとDateヘッダがあり、改行が2つ連続して存在するまでを　メールヘッダ、以降をメール本文として扱います。メール本文及びそれ以降が複数のパートで構成されている場合、後続のパート以降は添付ファイルとして扱います。

●添付ファイル展開

添付ファイルがMS-Office (Word / Excel / PowerPoint) / PDF ファイルなどの場合、デ変研TFライブラリと連動して、テキスト抽出後のファイルも取り出すことができます。

●圧縮ファイル展開

添付ファイルが圧縮ファイルであった場合に、圧縮ファイル内を展開して取り出しできます。添付ファイル中にメールファイルの添付や、圧縮内にメールファイルがあった場合、もしくは、添付ファイルの中にzip圧縮があり、そのなかにlha圧縮があるような階層圧縮ファイルも、展開して順に取り出すことができます。

●安全のための限界値設定

メモリ使用制限 (limit_total_memory)
１つのファイルの制限(limit_one_file)
ファイルの使用制限(limit_total_file)
ヘッダの最大値の指定(limit_eml_header)
メール本文の最大値の指定(limit_eml_body)
メールの入れ子展開の階層指定(limit_level)
テキストとして取り出す文字コードの指定 (language)

メールと圧縮ファイルの階層展開ライブラリ

デ変研 MFX ライブラリ

本ライブラリは、メールからヘッダ・本文・添付ファイルの情報を取り出す【メール展開(MX)】機能と、圧縮ファイルの展開をする【圧縮展開(FX)】機能を統合したライブラリです。
メールと圧縮ファイルを内部領域に展開し、１つ１つのメールの内部情報を取り出すことができます。
さらに mbox や PST(Outlook) といった複合メール形式にも対応しており、複合形式メールの最終個数を返した後、任意の１つのメールを取出し、展開することができます。

１つのメールの展開後の結果の表示「result.txt」ついて

１つのメール(eml形式、RFC822)の場合

メールヘッダ
メールボディ
１つ目の添付ファイル　(dehenken_word.doc)
２つ目の添付ファイル　(dehenken_xlsx.zip)

このzipファイル内には、dehenken_xlsx.xlsxを内包

上記１つのメールを本ライブラリが展開した場合、展開後のresult.txt の内容の参照方法

番号	TYPE	LEVEL	L_STRING	備考
No.0	E-MAIL	0	0	EMLファイル全体
No.1	EML HEADER	1	1	ヘッダ
No.2	EML BODY	1	2	本文
No.3	MS OFFICE	1	3	１つ目の添付
No.4	PKZIP	1	4	２つ目の添付
No.5	MS OFFICE VISTA XLSX	2	4.1	２つ目の添付の展開後の内容

■対応 OS

Red Hat Linux
AS3 / ES3 / WS3 /AS4 / ES4 / WS4 / EL5 / EL6 / EL7 / EL8
Windows
2000 / XP / Vista / 7 / 8 / 8.1 / 10
Windows Server
2000 / 2003 / 2008 / 2008R2 / 2012 / 2012R2 / 2016 / 2019
Windows Storage Server
2012R2 / 2016

■対応Ｃコンパイラ

Windows
Microsoft Visual Studio 2008以上
Linux
Gnu C Compiler(gcc)

■構成

メモリ　1GB以上
HDD利用量　500MB以上

※Windowsは、x86 また x64を対応に含めます。
※Linuxは、32bit版と64bit版の両方を対応に含みます。
※他のOS・コンパイラ・開発環境下でご不明な点は、お問い合わせください。
※ハードウェアの搭載メモリは推奨2GB以上で、メモリ量が多い方が大きな文書に対応できます。

result.txtの全内容

NUM : 6


```
*** No.0 ***
  TYPE      : E-MAIL
  DATA_SIZE : 43557
  TEXT_SIZE : 0
  LEVEL     : 0
  L_STRING  : 0
  NAME      : (null)
  STATUS    : 0

*** No.1 ***
  TYPE      : EML HEADER
  DATA_SIZE : 897
  DATA_ON_M :
  out_dir.2228_1456884341/MEMORY/data002
  TEXT_SIZE : 897
  TEXT_ON_M :
  out_dir.2228_1456884341/MEMORY/text002
  LEVEL     : 1
  L_STRING  : 1
  NAME      : (null)
  STATUS    : 0

*** No.2 ***
  TYPE      : EML BODY
  DATA_SIZE : 184
  DATA_ON_M :
  out_dir.2228_1456884341/MEMORY/data003
  TEXT_SIZE : 168
  TEXT_ON_M :
  out_dir.2228_1456884341/MEMORY/text003
  LEVEL     : 1
  L_STRING  : 2
  NAME      : (null)
  STATUS    : 0

*** No.3 ***
  TYPE      : MS OFFICE
  DATA_SIZE : 24064
  DATA_ON_M :
  out_dir.2228_1456884341/MEMORY/data004
  TEXT_SIZE : 186
  TEXT_ON_M :
  out_dir.2228_1456884341/MEMORY/text004
  LEVEL     : 1
  L_STRING  : 3
  NAME      : dehenken_word.doc
  STATUS    : 0

*** No.4 ***
  TYPE      : PKZIP
  DATA_SIZE : 6383
  DATA_ON_M :
  out_dir.2228_1456884341/MEMORY/data005
  TEXT_SIZE : 0
  LEVEL     : 1
  L_STRING  : 4
  NAME      : dehenken_xlsx.zip
  STATUS    : 0

*** No.5 ***
  TYPE      : MS OFFICE VISTA XLSX
  DATA_SIZE : 9363
  DATA_ON_M :
  out_dir.2228_1456884341/MEMORY/data006
  TEXT_SIZE : 162
  TEXT_ON_M :
  out_dir.2228_1456884341/MEMORY/text006
  LEVEL     : 2
  L_STRING  : 4.1
  NAME      : dehenken_xlsx.xlsx
  STATUS    : 0
```