デ変研監査ライブラリは、
次の３つのライブラリから構成されています。

デ変研TFライブラリ（テキスト抽出ライブラリ）

デ変研 TF ライブラリは、MS-Office、PDF、一太郎等の文書ファイルからテキストを抽出する、
組込型のテキストフィルタ（Text Filter）ライブラリです。
フォーマットの内部のバイナリデータを解析し、プロパティ情報とテキスト情報を抽出します。
その圧倒的な高速性は、全文検索におけるインデキシング生成時間の大幅な短縮化を実現します。

デ変研MFXライブラリ（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、メール内部の１つ１つのファイルの情報や、
圧縮ファイル内部の１つ１つのファイルの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つのメールを取り出し、展開することができます。

デ変研GGライブラリ
（個人情報を検出するための高速パターンマッチ技術ライブラリ）

デ変研GGライブラリは、辞書に登録された個人情報に基づいて、１つのファイル（バイナリ可）から、
さまざまな文字コードで書かれた個人情報や、マイナンバーやクレジットカード番号のような、
番号を検出することができるライブラリです。

■対応 OS

Red Hat Linux

AS3 / ES3 / WS3 / AS4 / ES4 / WS4 / EL5 /
EL6 / EL7 / EL8 / EL9 / EL10

Windows

2000 / XP / Vista / 7 / 8 / 8.1 / 10 /11

Windows Server

2000 / 2003 / 2008 / 2008R2 / 2012 / 2012R2 / 2016 /
2019 / 2022 / 2025

■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 または x64 を対応に含めます。
※Linuxは、x86,x64またはArm64を対応に含めます。
※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。

Dehenken

株式会社 データ変換研究所　Dehenken Limited

本　　社 〒600-8433 京都市下京区高辻通室町西入繁昌町 300-1 カノン室町四条 308
TEL 075-366-4645 FAX 075-366-4649
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 : https://www.dehenken.co.jp/ E-Mail : info_ml@dehenken.co.jp

EST'D 1999 Dehenken Limited © Copyright Dehenken 2025. All rights reserved.

品質マネジメントシステム ISO 9001:2015 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「ソフトウエア商品の営業、設計・開発・製造」です。

Certificate No. : C741343
Initial certification date : 27 September, 2011

個人情報ファイル検出エンジン

デ変研監査ライブラリ

AuditLibrary Dehenken

Dehenken

デ変研監査ライブラリは、ファイル中に存在する人名、住所、メールアドレス、郵便番号、電話番号、クレジットカード番号、マイナンバー、法人番号、ID番号を検出するライブラリです。Dehenkenの高速パターンマッチ技術（特許4152273号 パターン照合方法、パターン照合装置）を利用し、個人情報にあたる情報を高速に検出します。

日本語の漢字で記述されている辞書情報は、JIS・EUC・Shift_JIS・UTF-8・UTF-16（UTF-16LE / UTF-16BE）にソフトウェア内部で展開し、いずれかのコードで一致したことで検出とし、監査個数に数えます。
キーワードパターンマッチング技術により大量パターンから高速に検出（監査）を実現しています。使い方は、パターン検出のためのサンプルプログラムを用意しておりますので、プログラム経験のある方には容易に使い方をご理解いただけます。

●人名は日本人網羅率８０%以上に対応

人名は1000名字の辞書で日本人網羅率70%、2000名字の辞書では、日本人網羅率80%となります。

●対象ファイルはバイナリファイルにも対応

対象ファイルはバイナリファイルにも対応していますので、フォーマットの不明な場合でも、そのファイル中に、Shift_JIS や UTF-8 の文字コードで記述された人名や住所のデータが含まれていれば、個人情報の抽出は可能です。

●ユーザによる辞書の追加や作成が可能

検出のためのキーワード群を「辞書」といいます。この辞書はテキストファイルにより構成されます。この辞書はテキストファイルを加筆・変更・削除するだけで、ユーザによる辞書に変更ができます。あらかじめ用意されているサンプル辞書を参考に独自の辞書にて検出する単語を変更することができます。

個人情報の検出エンジン

デ変研監査ライブラリ

デ変研監査ライブラリは Dehenken の高速パターンマッチ技術を用いた個人情報検出エンジンです。対象ファイルは、バイナリやメール・圧縮されたファイル・MS-Office や PDF などの文書ファイルなど、あらゆるファイル（ただし暗号化されていないもの）の中から、個人情報を検出し数え上げることができます。

基本辞書

人名 (日本人名字の漢字・半角カナ)、地名 (都道府県市区町村)、
メールアドレス、電話番号（電話番号パターン）、
クレジットカード (数列パターン＆アルゴリズム)、
マイナンバー、会社番号の検出ができます。

Audit Library Dehenken

様々な文書フォーマットに対応

Microsoft Word / Excel / PowerPoint
Microsoft Visio / XPS / for Mac シリーズ
JustSystems 一太郎 / Adobe Systems Acrobat
PDF / RTF / テキスト文書 (EUC-JP / Shift_JIS / UTF-8)
HTML / XML / SGML / ODF (Writer / Calc / Impress)
バイナリファイルとのマッチング可

監査対象

●人名辞書

人名辞書は単漢字を除く名字の上位1000語と2000語の名字辞書と、単漢字名字辞書を提供しています。人名の監査においては、1000語と2000語の名字辞書により、日本人の名字の多いもの順にて構成していますので、日本全国の名字の網羅率は各70%と80%となります。

●住所辞書

住所の監査においては、47都道府県名と市区町村名を住所辞書として提供しています。市区町村名は、2004年度以降発行されている郵便番号表に基づくものを搭載しています。

●メールアドレス

メールアドレスの監査では辞書に依存せずに、監査対象データの「@」を基準として前後にメールアドレスとして許可されている文字列が続く範囲をメールアドレスとして検出します。

●郵便番号

辞書ファイルへ 3 桁 -4 桁または 7 桁の # によるパターンを記述することで郵便番号の検出に対応しています。

●電話番号

固定電話番号は、03 や 06 を先頭とし、####-#### を後続数値として記述することで検出できます。携帯電話番号は、辞書ファイルへ 050######## / 070######## / 080######## / 090########を記述することで検出できます。このように電話番号パターン辞書を14パターン用意しており、パターンとマッチするものを電話番号として取り出すことができます。

●マイナンバー（個人番号）

2015 年 10 月より施行されましたマイナンバー (個人番号) の監査では、辞書に依存せずに、数値列の桁数が 12 桁かつ政府が公開しているチェック方式にて識別された数値列を検出します。

●法人番号

2015 年 10 月より施行されました法人番号の監査では、辞書に依存せずに、数値列の桁数が 13 桁かつ政府が公開しているチェック方式にて識別された数値列を検出します。

●クレジットカード番号

クレジットカード番号の監査では、辞書に依存せずに、数列の桁数が14～16桁かつLuhnアルゴリズムによるチェックにて識別された数列を検出します。カードの桁数と検証アルゴリズムに差異があり、対応しているカード名は、次のようになります。

カード名	番号長	検証
VISA	16	Luhnアルゴリズム
MasterCard	16	Luhnアルゴリズム
DiscoverCard	16	Luhnアルゴリズム
JCB	16	Luhnアルゴリズム
Amex	15	Luhnアルゴリズム
DinersClub	14	Luhnアルゴリズム

次のカードは、Luhnアルゴリズムではないとされているので、検出対象から除外されます。

UTAP	15	1桁目1からはじまる15桁。アルゴリズムは不明。先頭番号不明。
中国銀聯	16	アルゴリズムの装着はなし。622126-622925, 624-626, 6282-6288

辞書ファイルに 15 / 16 桁の数列パターン及び 4 桁 -6 桁 -5 桁 / 4 桁 -4 桁 -4 桁 -4 桁の数列パターンを記述し、パターンとしてのみ検出させることで、対応は可能と考えます。

●ID番号

ASCII 文字列の任意の数字の羅列するパターンに対応をしています。
例、DHK-#### といったような ACII コードとの組み合わせパターンに対応しています。

判定基準ファイル

判定基準ファイルとは、個人情報の個々の対象基準について、最大の検出個数をあらかじめ登録しておくファイルです。その基準の示す最大の検出値に至るまで検出は継続します。その基準が満たされた時点で、これ以上の探索を継続しても、検出個数が増加することにはならないので、満たされて以降、検出処理を継続することはすべて無駄になります。したがって、検出量が満たされた場合には、速やかに判定処理に遷移します。このため情報検出基準と現在検出数は、常に比較する処理を行っています。
この判定基準を越えた場合、速やかに後処理に推移しますが、判定基準を超えなかった場合は、ファイルの最終位置まで到達した後で、後処理に推移します。後処理では、バンクによる監査が完了した後、判定基準値にしたがって判定を行います。