Hyper Estraier

http://bty.sakura.ne.jp/wp/archives/30

設置

```
/opt/hyperstraier にインストールする
```

■ QDBM のインストール

```
$ wget http://qdbm.sourceforge.net/qdbm-1.8.77.tar.gz
$ tar zxvf qdbm-1.8.77.tar.gz
$ cd qdbm-1.8.77
$ ./configure --prefix=/opt/hyperestraier/
$  make
#  make install
```

■ Hyper Estraier のインストール

PKG_CONFIG_PATH 環境変数を QDBM の pkg-config を指定する必要がある

```
export PKG_CONFIG_PATH=/opt/hyperestraier/lib/pkgconfig/
```

または、 configure 時にオプションで指定する

```
$ wget http://hyperestraier.sourceforge.net/hyperestraier-1.4.13.tar.gz
$ tar zxvf hyperestraier-1.4.13.tar.gz
$ cd hyperestraier-1.4.13
$ ./configure --prefix=/opt/hyperestraier PKG_CONFIG_PATH=/opt/hyperestraier/lib/pkgconfig/
$ make
$ make check
$ su
# make install
```

インデックスの作成

■インデックス作成

```
estcmd create インデックスファイル
```

でインデックスを作成する。create せずに gather でも作成できるが、属性などにもインデックスを指定したりする場合は create を使う。
また、gather で作成したインデックスに対して、後から属性をインデックスの対象にすることもできる。
例えば、ファイルサーバとして使う場合の例

```
estcmd create -attr _lpath str -attr _lreal str -attr @mdate seq -attr @size seq インデックス
```

規模が大きい倍は -xh 等のオプションを付ける。

■簡単な例

```
estcmd gather -il ja -sd インデックスファイル 対象のパス
```

例

```
estcmd gather -pc UTF-8 -il ja -sd casket /home/web/html
```

■インデックス作成のオススメオプション

```
estcmd gather -cl -il ja -sd -cm -pc UTF-8 -fx ".pdf"
"H@/opt/hyperestraier/share/hyperestraier/filter/estfxpdftohtml" -fx
".xls,.ppt,.doc,xlsx,pptx,docx,xlt" "T@/opt/hyperestraier/share/hyperestraier/filter/estfxmsototext"
/POI/estIndex ./sample
```

-cl
: 不要な領域を再利用

-il
: 言語指定。ja で日本語指定

-sd
: ファイルの更新日時を文書の属性として追加

-cm
: 文書の属性の更新日時がファイルの更新日時より古い場合にのみ登録を行う

-pc
: パスの文字コード

-cs
: 検索時に利用するメモリのキャッシュサイズをメガ単位で指定。デフォルトは 64M。規模が大きい場合は 1000 とか指定してもいいと思う。

-xh
: 規模が大きい場合は、-xh や -xh2 などを指定。

■ fx オプションについて

```
-fx
```

オプションは複数指定できる。また、 * で全てのファイルを対象にできる。
この時、複数の fx に合致する場合は、先に合致したもののみを処理する。
例えば

```
-fx ".txt" "T@ コマンド１ " -fx ".txt" "T@ コマンド２ " -fx "*" "T@ コマンド３ "
```

を指定した場合は、 txt 拡張子のファイルはコマンド１のみが実行される。

■一部の拡張子だけ再作成する場合

```
estcmd gather -cl -il ja -sd -pc UTF-8 -fx ".zip,.lzh"
"T@/opt/hyperestraier/share/hyperestraier/filter/estfxpathtotext" -fz /POI/estIndex ./sample
```

-fz
: -fx の条件に当てはまらないファイルを無視

-cm を外して、-fz を追加する。

-fr

```
-tf
```

は便利そうだけど、対象外のファイルのインデックスを削除してしまうので注意。

検索用ページの設置

・CGI の設置場所

- URL http://192.168.11.25/search/
- ローカルパス /home/web/html/search

```
$ cd /home/web/html/search/
$ cp /usr/local/libexec/estseek.cgi .
$ cp /usr/local/share/hyperestraier/estseek.* .
$ ls
estseek.cgi  estseek.conf  estseek.help  estseek.tmpl  estseek.top
```

- estseek.cgi CGI スクリプト
- estseek.conf 設定ファイル
- estseek.tmpl 検索ページのテンプレートファイル
- estseek.top 検索ページの初期画面のメッセージを記述したファイル
- estseek.help 検索機能の簡単な使い方を記述したファイル

・estseek.conf をエディタで修正する

```
indexname: /home/web/casket
...
replace: file:///home/web/html/{{!}}http://192.168.11.25/
...
```

・検索してみる

```
http://192.168.11.25/search/estseek.cgi
```

POI で MS ドキュメントを検索する

■ POI のダウンロード

http://poi.apache.org/download.html

■ MS ドキュメント解析プログラム作成

MSOfficeToText.java

■コンパイル

```
poi-ooxml-3.9-20121203.jar
poi-3.9-20121203.jar
poi-scratchpad-3.9-20121203.jar
ooxml-lib/dom4j-1.6.1.jar
ooxml-lib/xmlbeans-2.3.0.jar
poi-ooxml-schemas-3.9-20121203.jar
```

あたりにクラスパスを通して

```
javac MSOfficeToText.java
```

■ share/hyperestraier/filter にシェル追加

■インデックス作成

```
   estcmd gather -cl -il ja -sd -cm -pc UTF-8 -fx ".pdf"
"H@/opt/hyperestraier/share/hyperestraier/filter/estfxpdftohtml" -fx
".xls,.ppt,.doc,xlsx,pptx,docx,xlt"  "T@/opt/hyperestraier/share/hyperestraier/filter/estfxmsototext"
-tr  /POI/estIndex ./sample
```

ファイルのパスを内容として登録する

アーカイブファイルとかバイナリファイルとかをせめてパスで検索でヒットさせたい場合、簡単なフィルタを作成してパスを登録するようにする。
share/hyperestraier/filter に移動して

```
cp estfxpdftohtml estfxpathtotext
vi estfxpathtotext
```

で、変数とか色々修正する。

```
#output the result
```

の下の行を

```
echo "$inputfile" | output
```

みたいな感じにする。

以下整理中

xlsx、docx、pptx を検索対象にする

http://ru11en.wordpress.com/category/linux/

```
.docx の場合 /usr/bin/unzip -caq "$infile" */document.xml
.xlsx の場合 /usr/bin/unzip -caq "$infile" */sharedStrings.xml */drawing1.xml
.docx の場合 /usr/bin/unzip -caq "$infile" */slide[0-9]*.xml
```

xlsx の drawing1.xml はテキストオブジェクト等のオブジェクト内のテキストが含まれている。
もし、xlsx 内にオブジェクトがない場合は、drawing1.xml は存在しない。
その場合は、ファイルが見つからないと出るが処理的には問題ないはず。

■インデックス作成コマンド

例

```
   estcmd     gather     -cl     -il     ja     -sd     -cm     -pc     UTF-8     -fx     ".pdf"
"H@/usr/local/share/hyperestraier/filter/estfxpdftohtml" ~/estIndex ~/temp2
   estcmd     gather     -cl     -il     ja     -sd     -cm     -pc     UTF-8     -fx     ".pdf"
"H@/usr/local/share/hyperestraier/filter/estfxpdftohtml"                -fx                ".xlsx"
"H@/usr/local/share/hyperestraier/filter/estfx_ooxml2xml.sh" -tr ~/estIndex ~/temp2
```

AD と連携して、権限によって検索結果を切り替える

調査中

http://list-archives.org/2013_/03_/26_/fess-user-lists-sourceforge-jp/fess-user-684_-%09_fess-%E3_%81%AE-active-directory-%E8%AA%8D%E8%A8%BC%E3%82%92%E5%85%83%E3%81%AB%E3%83%AD%E3%83%BC%E3%83%AB%E3%83%99%E3%83%BC%E3%82%B9%E6%A4%9C%E7%B4%A2/f/5017227789

権限によって実行するプログラムを変更する

http://www.tamashima.biz/archives/2009/09/hyperestraier_1.html

ユーザの権限によって実行結果を変える方法として、
estseek.cgi、estseek.conf をコピーして実行権限を変更しまう方法もある。

Fess

デスクトップ検索の設定

http://fess.sourceforge.jp/ja/4.0/config/desktop-search.html

デスクトップ検索機能はデフォルトでは無効になっています。 以下の設定により有効にしてください。
まず、bin/setenv.bat を以下のように java.awt.headless を true から false に編集します。

```
... -Djava.awt.headless=false ...
```

次に、webapps/fess/WEB-INF/conf/crawler.properties に以下を追加します。

```
search.desktop=true
```

上記を設定したら、Fess を起動してください。基本的な利用方法は特に変わりません。

Solr

今のところ使う予定なし。
ちょっと動かして気がついたことを書く。

起動時にポートを変える

http://ameblo.jp/itboy/entry-11521145965.html

```
etc/jetty.xml
```

の

```
<Set name="port"><SystemProperty name="jetty.port" default="8983"/></Set>
```

を

```
<Set name="port"><SystemProperty name="jetty.port" default="8081"/></Set>
```

とか。

ポートを変更したサーバに対して、update を行う場合は

```
java -Durl="http://localhost:8081/solr/update" -jar post.jar *.xml
```

とか。