

# 全文検索システム『ひまわり』/簡単な検索用データの作成方法

[Top](#) / [全文検索システム『ひまわり』](#) / 簡単な検索用データの作成方法

言語を選択 | ▼

## 全文検索システム『ひまわり』

### 目次


- [1. はじめに](#)
- [2. 用意するもの](#)
- [3. 作成手順](#)
  - [3.1 書誌情報のタグ付け](#)
  - [3.2 複数の文書を一度に検索できるようにする](#)
  - [3.3 文書の保存](#)
  - [3.4 作成した文書のインストール](#)
  - [3.5 索引付け](#)

## 1. はじめに <sup>↑</sup>

- 『ひまわり』の検索用データの作成方法について説明します。
- ここでは、タグ付けされていないテキストに書誌情報をタグ付けした簡単な XML 文書を作成します。
- 作業環境として、Windows 環境を想定しています。

<sup>↑</sup>

## 2. 用意するもの <sup>↑</sup>

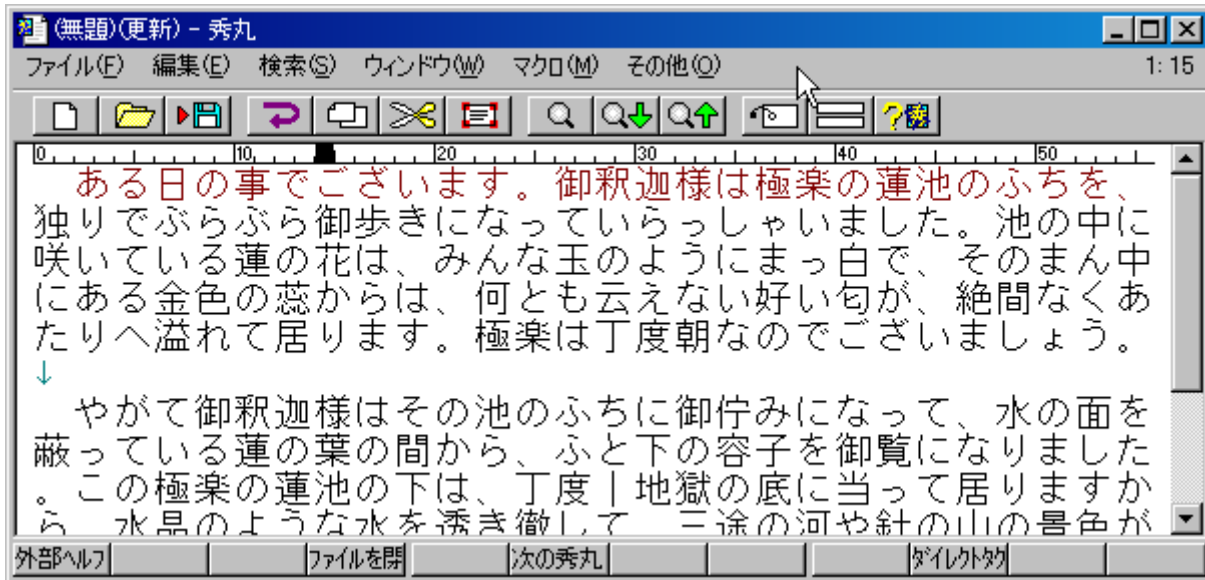
- [『ひまわり』\(ver.1.3以降\)](#) のインストール
- 検索対象のテキスト
- [「秀丸エディタ」](#)
  - なお、「秀丸エディタ」以外のエディタをお使いの方は、次の条件のファイルを作成できるエディタをご用意ください。お使いのエディタで作成できない場合は、文字コード変換プログラムを利用して、文字コードと改行コードを変換してください。
    - 文字コード：Unicode (UTF-16 Byte Order Mark 付き)
    - 改行コード：LF
-  [『ひまわり』用設定ファイル simpledoc.zip](#) ... ダウンロードしておいてください。

<sup>↑</sup>

## 3. 作成手順 <sup>↑</sup>

### 3.1 書誌情報のタグ付け <sup>↑</sup>

検索対象のテキストをエディタで開いてください。ここでは、芥川龍之介の「蜘蛛の糸」を開いています。



次に書誌情報として、「著者」と「タイトル」を付与することにします。まず、文書の先頭に次のタグを付け加えてください。これを「開始タグ」と言います。

```
<simplifiedoc タイトル="蜘蛛の糸" 著者="芥川龍之介">
```

上の例のように、半角の <> で囲われた部分がタグです。「simplifiedoc」は、タグの名前です。このタグの属性として、「タイトル」と「著者」を埋め込みます。= や " は半角であることに注意してください。

```
<simplifiedoc タイトル="蜘蛛の糸" 著者="芥川龍之介">
```

```
ある日の事でございます。御釈迦様は極楽の蓮池のふちを、独りでぶらぶら御歩き
になっていらっしゃいました。池の中に咲いている蓮の花は、みんな玉のようにまっ
白で、そのまん中にある金色の蕊からは、何とも云えない好い匂が、絶間なくあたり
へ溢れて居ります。極楽は丁度朝なのでございましょう。
```

次に、文書の末尾に、開始タグと対応する「終了タグ」の「</simplifiedoc>」をつけます。これで、開始タグと終了タグで囲まれた範囲の書誌情報が記述できたことになります。

```
しかし極楽の蓮池の蓮は、少しもそんな事には頓着致しません。その玉のような白
い花は、御釈迦様の御足のまわりに、ゆらゆら萼を動かして、そのまん中にある金色
の蕊からは、何とも云えない好い匂が、絶間なくあたりへ溢れて居ります。極楽もも
う午に近くなったのでございましょう。
</simplifiedoc>
```

最後に、今作成した文書全体を「corpus」タグで囲います。文書の先頭には、開始タグの「<corpus>」、文書の末尾には、終了タグの「</corpus>」をつけてください。

```
<corpus>
```

```
<simplifiedoc タイトル="蜘蛛の糸" 著者="芥川龍之介">
```

```
ある日の事でございます。御釈迦様は極楽の蓮池のふちを、独りでぶらぶら御歩き
になっていらっしゃいました。池の中に咲いている蓮の花は、みんな玉のようにまっ
白で、そのまん中にある金色の蕊からは、何とも云えない好い匂が、絶間なくあたり
へ溢れて居ります。極楽は丁度朝なのでございましょう。
```

```
：（中略）
```

```
しかし極楽の蓮池の蓮は、少しもそんな事には頓着致しません。その玉のような白
い花は、御釈迦様の御足のまわりに、ゆらゆら萼を動かして、そのまん中にある金色
の蕊からは、何とも云えない好い匂が、絶間なくあたりへ溢れて居ります。極楽もも
う午に近くなったのでございましょう。
```

```
</simplifiedoc>
```

```
</corpus>
```

以上で、タグ付けは終了です。



## 3.2 複数の文書を一度に検索できるようにする <sup>↑</sup>

3.1 では、一つの作品に対して、書誌情報をつけました。しかし、たくさんの作品を一度に検索したいことがよくあると思います。

そこで、別の作品を追加する方法について説明します。ここでは、同じ芥川龍之介の「猿蟹合戦」を追加してみます。なお、一つの文書を検索できるだけでよい場合は、この節は読み飛ばしてかまいません。

追加方法は簡単で、「蜘蛛の糸」のあとに、「猿蟹合戦」を追加するだけです。エディタで追加するテキストをコピーして、「蜘蛛の糸」のあとに貼り付けましょう。結果は、次のようになります。

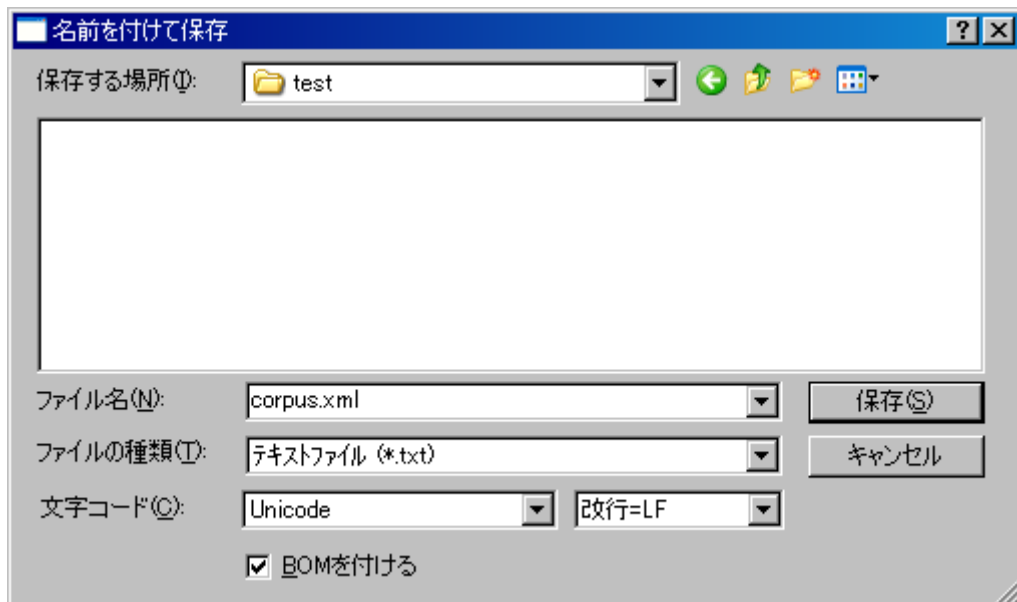
```
<corpus>
<simplifiedoc タイトル="蜘蛛の糸" 著者="芥川龍之介">
  ある日の事でございます。御釈迦様は極楽の蓮池のふちを、独りでぶらぶら御歩き
  : (中略)
  の蕊からは、何とも云えない好い匂が、絶間なくあたりへ溢れて居ります。極楽ももう午に近くなったのでございましょう。
</simplifiedoc>
<simplifiedoc タイトル="猿蟹合戦" 著者="芥川龍之介">
  蟹の握り飯を奪った猿はとうとう蟹に仇を取られた。蟹は曰、蜂、卵と共に、怨敵の猿を殺したのである。——その話はいまさらしないでも好い。ただ猿を仕止めた後、蟹を始め同志のものはどう云う運命に逢着したか、それを話すことは必要である。なぜと云えばお伽噺は全然このことは話していない。
  : (中略)
  とにかく猿と戦ったが最後、蟹は必ず天下のために殺されることだけは事実である。語を天下の読者に寄す。君たちもたいてい蟹なんですよ。
</simplifiedoc>
</corpus>
```

さらに別の文書を追加したい場合も、同じ方法で追加することができます。ただし、追加した結果の文書全体を「corpus」タグで囲うのを忘れないでください。

### 3.3 文書の保存 [↑](#)

次に、作成した文書を保存します。保存するときのファイル名は、corpus.xml としてください。また、すでに、説明したように、文字コードはUnicode (UTF-16, Byte Order Mark 付き)、改行コードは LF としてください。


「秀丸エディタ」では、次の設定で保存します。「BOMをつける」がチェックされていることに注意してください。



### 3.4 作成した文書のインストール [↑](#)

次に、作成した文書(corpus.xml)を『ひまわり』にインストールします。『ひまわり』がまだインストールされていない場合は、[『ひまわり』利用者マニュアル](#)を参照して、『ひまわり』のインストールを完了させてください。最新版の『ひまわり』は[ダウンロードのページ](#)にあります。

corpus.xml のインストールは、次の手順で行ってください。

1. 「準備」のところで示した  [simplifiedoc.zip](#) を解凍してください。このファイルは、zip 形式で圧縮されています。Windows であれば、マウスでファイルを右クリック後、「すべて展開」を行うことにより、解凍できます。
2. 解凍すると、「Corpora」というフォルダがあるはずです、このフォルダを『ひまわり』がインストールされているフォルダに移動してください。
3. 「Corpora」フォルダの中に、「Simplifiedoc」フォルダがあるはずです。このフォルダに、corpus.xml を移動してください。
4. 「Corpora」フォルダの中に、設定ファイルの「config\_simplifiedoc.xml」があるはずです。このファイルを、『ひまわり』がインストールされているフォルダにコピーしてください。

### 3.5 索引付け

次に、作成した corpus.xml に対して、「索引付け」を行います。「索引付け」は、高速に全文検索するために必要な処理です。索引付けの手順は、次のとおりです。

1. 『ひまわり』を起動してください。
2. [ファイル]→[新規]を実行し、config\_simplifiedoc.xml を読み込んでください。
3. [ツール]→[インデックス生成]を実行してください。  
**注：**索引付けを再度行う場合は、すでに作成されている索引ファイルをすべて削除してください。索引ファイルは、「Simplifiedoc」フォルダ中の拡張子が .cix, .eix, .aix のファイルです(例：corpus.sd.cix)。
4. 「インデックス生成が終了しました。」と表示されれば、索引付けは終了です。

以上で、検索用データ作成は終了です。実際に検索してみてください。

---

Last-modified: 2012-02-07 (火) 13:33:53 (2397d)

Site admin: [anonymous](#)

**PukiWiki 1.4.7** Copyright © 2001-2006 [PukiWiki Developers Team](#). License is [GPL](#).  
Based on "PukiWiki" 1.3 by [yu-ji](#). Powered by PHP 5.1.6. HTML convert time: 0.104 sec.