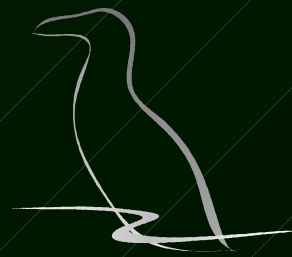
Lightweight Language Day and Night

The updates of the Xgawk

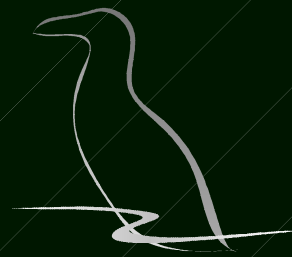
Hirofumi Saito <hi_saito@yk.rim.or.jp>

Xgawk Developers Group in SF.net
https://sourceforge.net/projects/xmlgawk/



What's AWK?

AWK って何?



意外に知られていない AWK

一昨年の「LL Satuaday」のアンケートにて、AWK を知っている人が非常に少ないことが判明・・・ちょっと意外 (残念)。

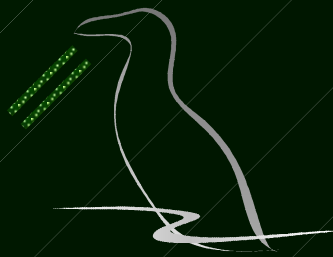
AWK とは 1977 年に Aho, Weinberger, Kernihan の 3 名が作成したテキスト処理に特化した言語で、以前は grep, sed と並んで UNIX の「三種の神器」とまで呼ばれた。特徴は以下のとおり。

- 変数に $ や @ が付かず、C 言語に近い文法
- 行 (レコード) をフィールドに自動的に分割
- 非常に小さい (ソースコードで 2 MB、Debian の実行ファイルで 300 KB 程度)

現在は GNU 版 gawk を使用するケースが多く、現 gawk メンテナーの Arnold Robbins 曰く、

「AWK は PERL に似た言語であるが、ちょっとばかし洗練されている!」

だそうな。



[おまけ]意外に知られていない Auk

AWK のマスコットは同じ発音の Auk (ウミツバメ) です。Perl のラクダなんかと比べると知らない人の多いウミツバメですが、こんな鳥です。

「Effective AWK Programming」の表紙よりも可愛いかも。

AWK ってどんなもの?

AWK は標準入力を含めたファイルの入出力に特化している関係で、ファイルを読む前段取りである BEGIN ブロック、ファイルを読み込むアクション、ファイルを読んだ後の END ブロックの 3 つのブロックだけで構成される。

```
BEGIN {
  ファイルを読む前の処理
}
/Pattern/ {
  正規表現 Pattern にマッチする時の処理
}
END {
  ファイルを読んだ後の処理
}
```

分かりやすいでしょ!

- 命令数も 40 くらい (使うのは 10 くらい) しかないので、簡単に覚えられる。
- 使用メモリも最低 3 MB 弱。

まさに人と資源に優しい LL。

国際化状況

gawk は国際化が行われています。現在、gawk 3.1.5 がリリースされ、length(), substr(), index() 等の関数でもバイト単位ではなくキャラクター単位になった。

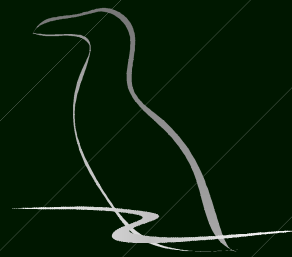
```
$ gawk --version | head -1
GNU Awk 3.1.4
$ LANG=ja_JP.EUC-JP gawk 'BEGIN{print length("あいうえお")}'
10
$ LANG=C gawk 'BEGIN{print length("あいうえお")}'
10

$ gawk --version | head -1
GNU Awk 3.1.5
$ LANG=ja_JP.EUC-JP gawk 'BEGIN{print length("あいうえお")}'
5
$ LANG=C gawk 'BEGIN{print length("あいうえお")}'
10
```

Multi-byte Extension 版が最新の gawk に追従できていなかったものの、ようやく日本語を自由に使える gawk がリリースされた。

What's Xgawk?

Xgawk って何?



先行開発の gawk、それが Xgawk

Xgawk は、Jürgen Kahrs 氏 (今日までしばらく音信普通でした) らが中心となり、Arnold Robbins 氏が gawk に今後組み込もうと考えていることを先行して開発。

- XML 対応 (node 単位でフィールドを扱える)
- XML に関して入出力は UTF-8 だけでなく、EUC-JP も使用可能
- PostgreSQL データベースの読み込み (MySQL も検討中)
- gawk からのフォークではなく、先行開発
- Arnold Robbins 氏好みの patch を提供 (苦笑)

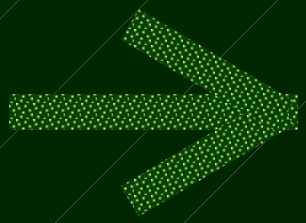
以前、xmlgawk という名称で開発をスタートしたが、現在 Xgawk と名称を変更。
本家 gawk 3.1.5 に追従予定。(未了)

Xgawk と gawk は "Fedora Core" と "Red Hat Enterprise Linux" の関係に似ているかも!? (決して不安定という意味ではないです・・・)

(gawk 開発環境は Fedora Core 4 だそうです)

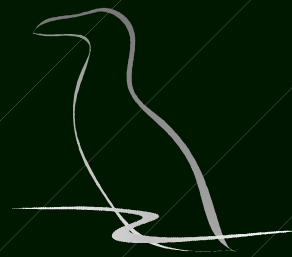
Gentoo Linux に採用されてますが・・・

Gentoo Linux に xmlgawk は採用されているものの、一部の人には不人気で「/usr/bin に入っていて、AWK はコアなプログラムのひとつなので、怪しいものは入れない」という意見もある。



xmlgawk の開発スタンスとしては、ビルド時に configure で指定しない限り、さらに XML mode に切り替えない限りは、普通の gawk として働くようにしてある。

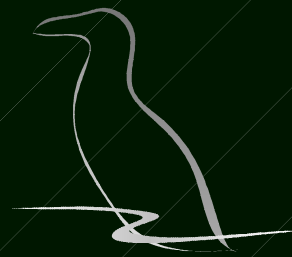
現時点では、Gentoo Linux の BTS をトレースできているわけではないが、問題は聞こえていない。



簡単な例の紹介 (例文)

以下のような日本語の XML も扱うことができます。

```
<?xml version="1.0" encoding="utf-8"?>
<!-- サンプルファイル(sample file) -->
<本-book>
  <本の情報-bookinfo id="こんにちは-hello-world">
    <タイトル-title>こんにちは-Hello, world</タイトル-title>
  </本の情報-bookinfo>
  <章-chapter id="序章-introduction">
    <タイトル-title>序章-Introduction</タイトル-title>
    <段落-para>これは序章です。ふたつの章があります。
                -This is the introduction. It has two sections</段落-para>
    <項目-sect1 id="この本について-about-this-book">
      <タイトル-title>この本について-About this book</タイトル-title>
      <段落-para>これは私が最初に書いた DocBook のファイルです。
                  -This is my first DocBook file.</段落-para>
    </項目-sect1>
    <項目-sect1 id="作業中-work-in-progress">
      <タイトル-title>注意-Warning</タイトル-title>
      <段落-para>まだ準備中です。
                 -This is still under construction.</段落-para>
    </項目-sect1>
  </章-chapter>
</本-book>
```



簡単な例の紹介

例 node の数を調べるには以下のようなスクリプトを用いる。

```
BEGIN { XMLMODE=1; nodes=0; XMLCHARSET="UTF-8"}
XMLSTARTELEM { nodes ++ }
END { print nodes }
```

→ 12

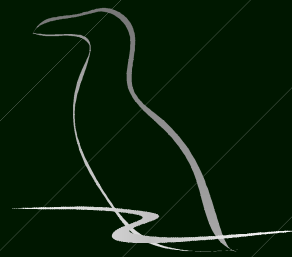
例 node の深さを調べるには以下のようなスクリプトを用いる。

```
BEGIN { XMLMODE=1; depth=0; XMLCHARSET="UTF-8"}
  XMLSTARTELEM {
    depth++
    if (depth > max_depth)
      max_depth = depth
    }
  XMLENDELEM   { depth-- }
END { print max_depth }
```

→ 4

ちょっと予約語が多く、慣れるまでに時間がかかりそう。

(CVS からのビルドも結構大変)



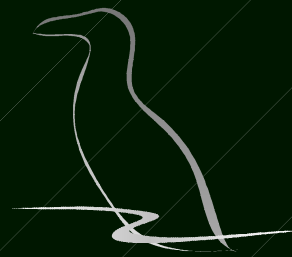
今後の gawk

gawk 3.1.5 では XML 対応は行われなかったものの、TCP などでインターネットのファイルを扱えるため、以下のような利点が考えられ、今後を期待したい。

- Web ページ上の HTML ファイルの処理
- RSS などの XML ファイルの簡易的な処理

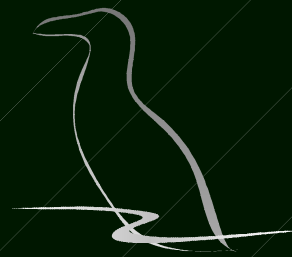
現在、xmlgawk プロジェクトでは開発者のみならず、テスターやドキュメントライターも募集しているので、興味のある人は参加してみてください。

Xgawk Developers Group in SF.net
https://sourceforge.net/projects/xmlgawk/



Message from Arnold

gawk メンテナー Arnold Robbins 氏からのメッセージ



Arnold Robbins 氏からのメッセージ (日本語)

こんにちは。awk についての手紙を依頼してくれてありがとうございます。この言語を自分のものにすることができればうれしいのですが、それは本当にそれを作成したベル研究所で聡明な方々のものになります。

awk はしばしば「小さい」言語と呼ばれます。特徴の点では、それは本当です。しかし、それが単独と組み合わせで持つ特徴は、パワーとエレガントの両方を提供してくれます。小さく、読みやすい (そして、書くのも簡単です !) プログラムで多くのことができます。awk を学ぶ時間を費やすなら、ずっと多くの時間を他の言語のいくつかを学びながら費やす必要はないと本当にわかるでしょう。

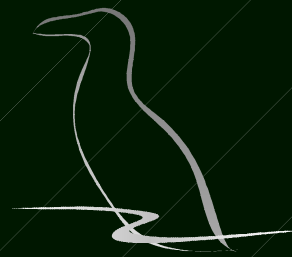
敬具

Arnold Robbins

時間がないので、翻訳したものを用意しました。

Wiki using AWK

AWK でできた Wiki



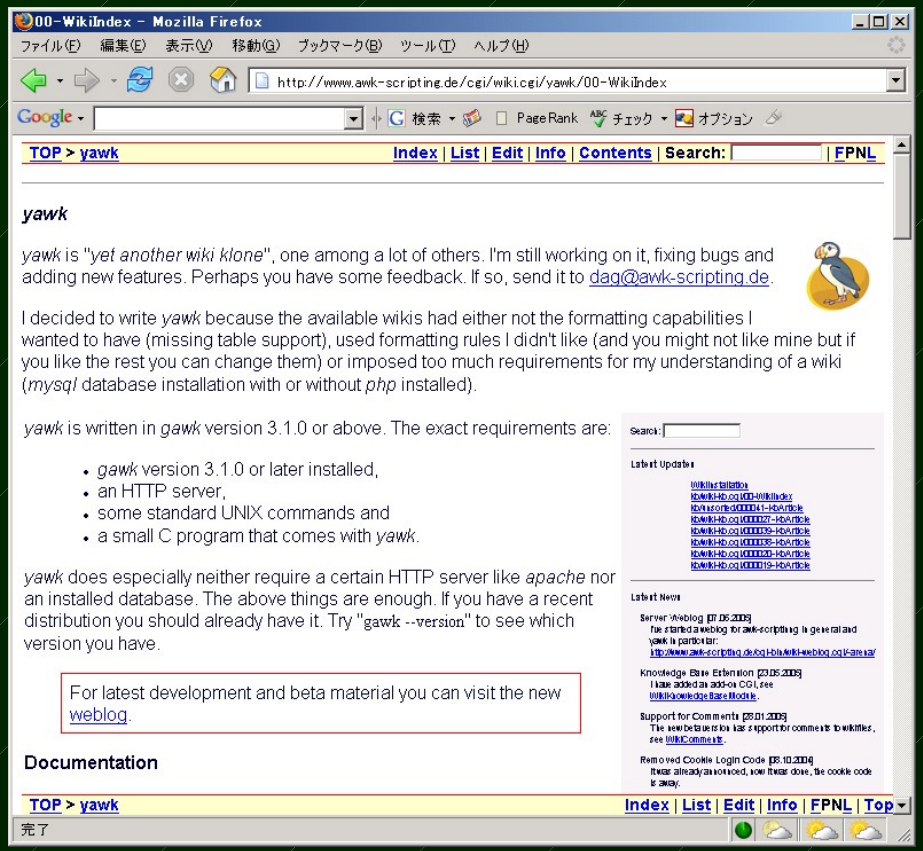
yawk

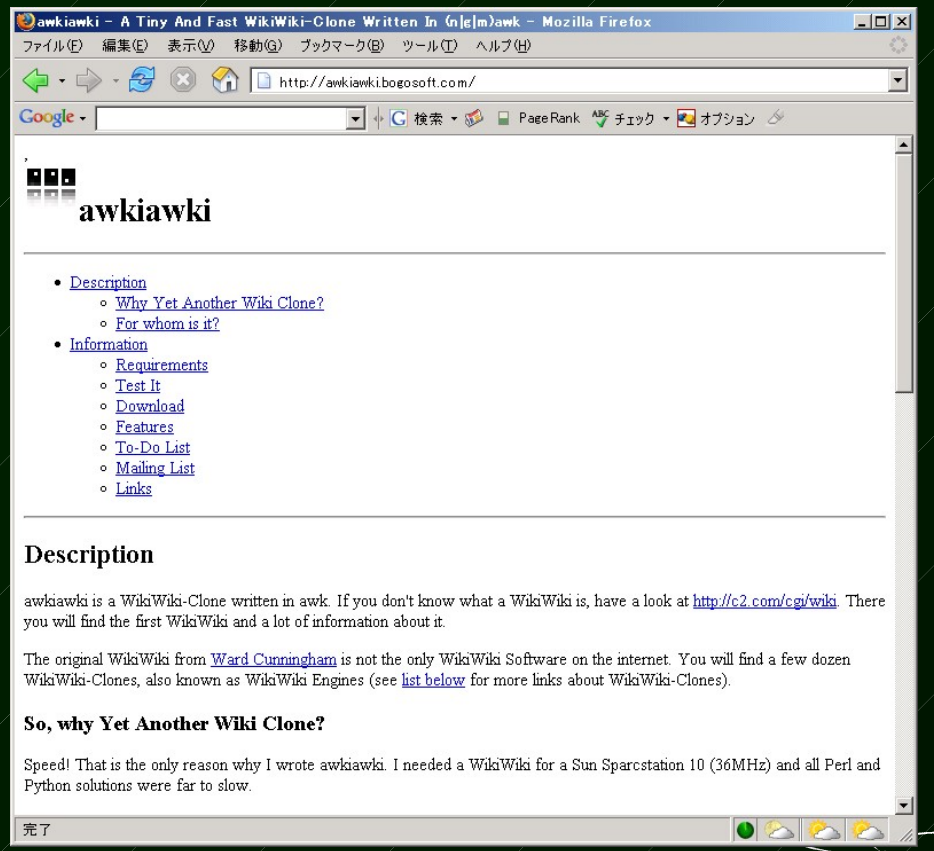
http://www.awk-scripting.de/cgi/wiki.cgi/yawk/00-WikiIndex

awkiawki

http://awkiawki.bogosoft.com/

awk でできた Wiki たち。今後の展開を期待。(来年はフレームワークとして参加??)





Fin.

ご清聴ありがとうございました