

Programme d'extraction des balises DESCRIPTION dans les fils RSS

Pour commencer, le texte du programme que vous pourrez réutiliser *en le modifiant* dans votre programme personnel pour traiter l'arborescence complète de tous les fils RSS du Monde pour l'année 2008 :

```

1  #!/usr/bin/perl
2  use XML::RSS;
3  my $file="$ARGV[0]";
4  my $rss=new XML::RSS;
5  #-----
6  my $TEXTEDUFIL="";
7  $TEXTEDUFIL."<FILE=\"\$file\">\n";
8  #-----
9  $rss->parsefile($file);
10 my $nombredechampdescription=0;
11 foreach my $item (@{$rss->{'items'}}) {
12     $nombredechampdescription++;
13     $TEXTEDUFIL."<article=\"\$nombredechampdescription\">\n";
14     my $description=$item->{'description'};
15     $TEXTEDUFIL."&#160; ". $description. "\n";
16 }
17 #-----
18 print $TEXTEDUFIL;
19

```

Ce programme est lancé de la manière suivante :

```
perl extract-txt-avec-xml-rss.pl 0,2-3208,1-0,0.xml
```

le programme prend en entrée un fil RSS passé en argument au programme dans la ligne de commandes.

Contrairement à ce que nous avons fait dans les séances précédentes (lecture du fil ligne à ligne et repérage par des expressions régulières du contenu de la balise description de chaque item du fil RSS), ce programme utilise une bibliothèque nommée XML::RSS (cf ligne 2) dont l'objectif est justement de manipuler « facilement » ce type de fichier pour accéder aux « différentes zones » définies par le format RSS utilisé par ce fichier.

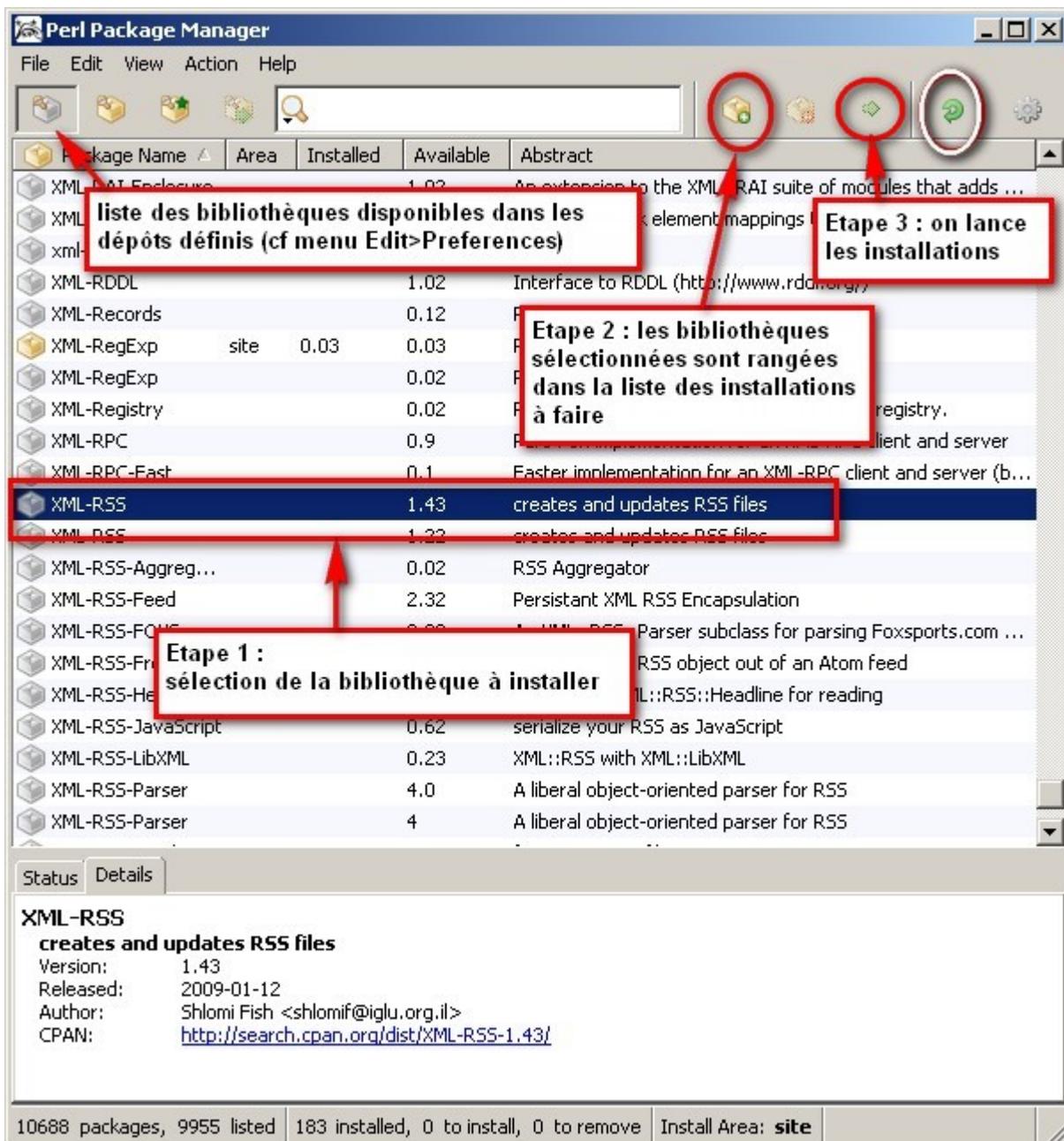
(cf <http://www.perl.com/pub/a/2000/01/rss.html>)

Cette bibliothèque n'est pas installée par défaut. Pour tester ce programme, il faut donc au préalable l'installer.

La figure qui suit montre comment réaliser cette installation pour la version de Perl ActiveState installée sur les machines du labo à l'ILPGA : (Perl ActiveState 5.8 pour windows). Le principe est le même pour toutes les versions Perl ActiveState (macosx...).

Pour installer une (ou plusieurs bibliothèques), on va « se servir » dans les dépôts (de bibliothèques) disponibles en ligne (sur le site d'ActiveState ou ailleurs).

Pour faire « ce marché », on utilise le programme « *Perl Package Manager* » (qui fait partie de la distribution Perl ActiveState) et qui est en fait le gestionnaire des bibliothèques pour Perl. PPM donne à voir les bibliothèques installées sur notre machine de travail et celles disponibles dans les différents dépôts disponibles (cf menu Edit>Preferences). Dans la figure qui suit, PPM est lancé (menu démarrer>ActivePerl), on sélectionne la bibliothèque à installer, puis on l'installe...



On peut ensuite l'utiliser...

```

bash-2.02$ perl extract-txt-avec-xml-rss.pl 0,2-3208,1-0,0.xml
<FILE="0,2-3208,1-0,0.xml">
<article="1">
  2 Le pape a exprimé sa "solidarité totale et incontestable" avec le peuple juif
  lors de son audience générale hebdomadaire, mercredi. Quelques jours après sa dé
  cision de lever l'excommunication de quatre évêques intégristes, parmi lesquels
 Mgr Williamson, auteur de propos négationnistes.<img width='1' height='1' src='h
  ttp://rss.feedsportal.com/c/205/f/3050/s/2e7d602/mf.gif' border='0' /><br/><br/>
  a href="http://da.feedsportal.com/r/28878066037/u/159/f/3050/c/205/s/48748034/a2
  .htm"></a>
<article="2">
  2 Au moment où tout le monde semble s'inquiéter, surtout dans le secteur automob
  ile, voilà quelqu'un qui n'est pas du tout morose. Optimiste, passionnant, enthousiasme,
  rusé, nous ne sommes pas vraiment dans le champ lexical de la crise avec Jean-Dominique
  Wagret, vice-président du pôle de compétitivité pour l'automobile ...<img width='1' height='1' src='http://rss.feedsportal.com/c/205/f/3050/s/2e7d603/mf.gif' border='0' /><br/><br/>
  a href="http://da.feedsportal.com/r/28878066036/u/159/f/3050/c/205/s/48748035/a2.htm"></a>
<article="3">
  2 Qui est le meilleur pote de Jérôme Fernandez en équipe de France ? Quel joueur
  étranger aimerait-il voir devenir français pour pouvoir jouer avec lui chez les
  Bleus ? L'hotellier d'Osijek était-il plus confortable que celui de Zagreb ? Qui est
  le joueur le plus drôle de l'équipe ? La France va-t-elle devenir championne d'
  Europe...<img width='1' height='1' src='http://rss.feedsportal.com/c/205/f/3050/s/2e7d604/mf.gif' border='0' /><br/><br/>
  a href="http://da.feedsportal.com/r/28878066035/u/159/f/3050/c/205/s/48748036/a2.htm"></a>
<article="4">
  2 Le constructeur aéronautique américain a annoncé, mercredi 28 janvier, la suppression
  cette année de 10 000 emplois, soit une réduction de 6 % de ses effectifs.<img width='1' height='1' src='http://rss.feedsportal.com/c/205/f/3050/s/2e7d605/mf.gif' border='0' /><br/><br/>
  a href="http://da.feedsportal.com/r/28878065195/u/159/f/3050/c/205/s/48745855/a2.htm"></a>
<article="5">

```

Comment ça marche...

le fonctionnement est décrit sur les sites suivants :

- <http://www.perl.com/pub/a/2000/01/rss.html>
- <http://search.cpan.org/~shlomif/XML-RSS-1.43/lib/XML/RSS.pm#DESCRIPTION>
- <http://perl-rss.sourceforge.net/>

On en parle au cours de la prochaine séance...