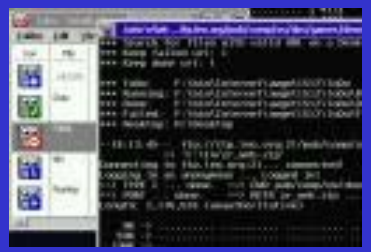



# *Programmation et projet encadré*

## *Boîte à outils* préambule





# Sommaire

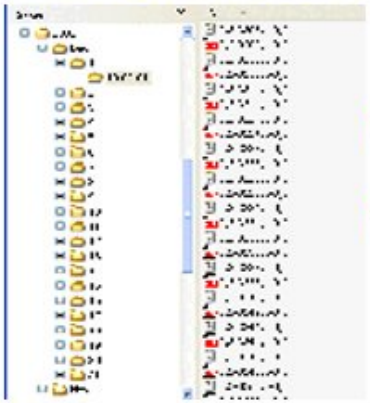
- Les séances « BOITES A OUTILS »
  - *Série 1* : Perl (filtrage, nettoyage…)
  - *Série 2* : Etiquetage (*Treetagger*, *Cordial*)
  - *Série 3* : Extraction terminologique
  - *Série 4* : Des textes aux Graphes (*via* Pajek)
    - Les mots qui s'attirent dans les fils (information mutuelle)

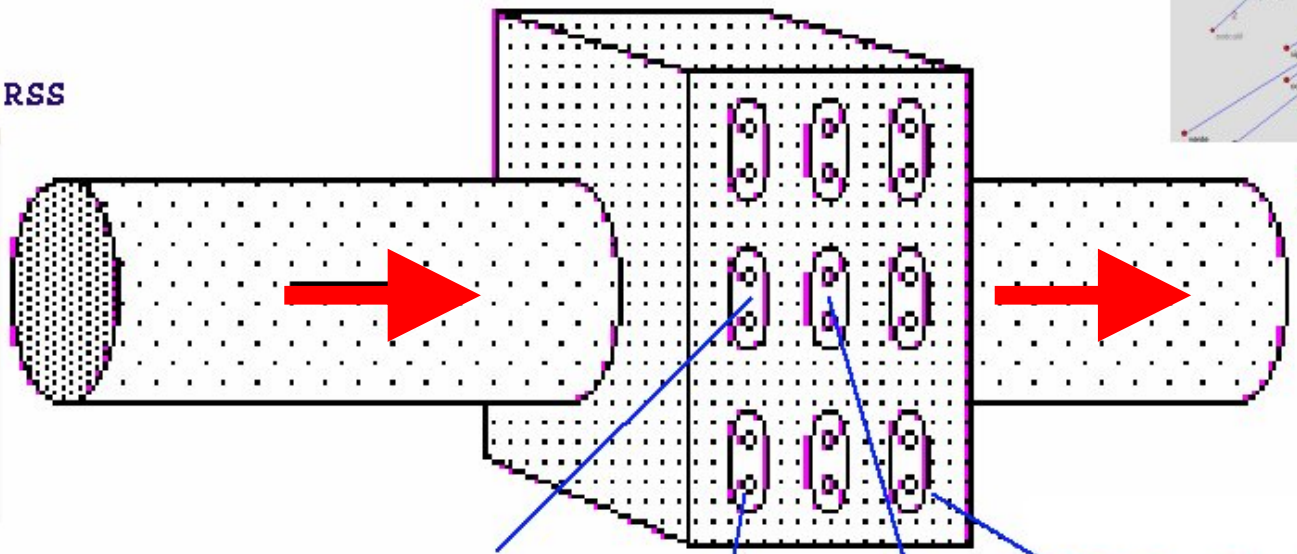
# Cartographie du « projet BàO »

progression

**IN**

un corpus de fils RSS





Des graphes de mots qui s'attirent

**OUT**

Boîtes à Outils

BàO série 1 — *Outil : perl, script de filtrage, script de nettoyage*

BàO série 2 — *Outils : treetagger, cordial et perl comme "glue"*

BàO série 3 — *Outils : perl, script de filtrage de patrons*

BàO série 4 — *Outils : Pajek, outils XML, script perl...*

# Corpus de travail

- Fils RSS du journal *Le Monde*
  - 17 fils RSS archivés une fois par jour (19h00) sur plusieurs semaines
  - Chacun des fils est accompagné de sa version « textuelle » (dite *profonde*) au format Lexico3
    - Le texte du fil + le texte complet de l'article associé au fil
    - Ces fichiers ne seront pas utilisés dans les BàO…
- Période traitée :
  - 20/11/2006-21/12/2006

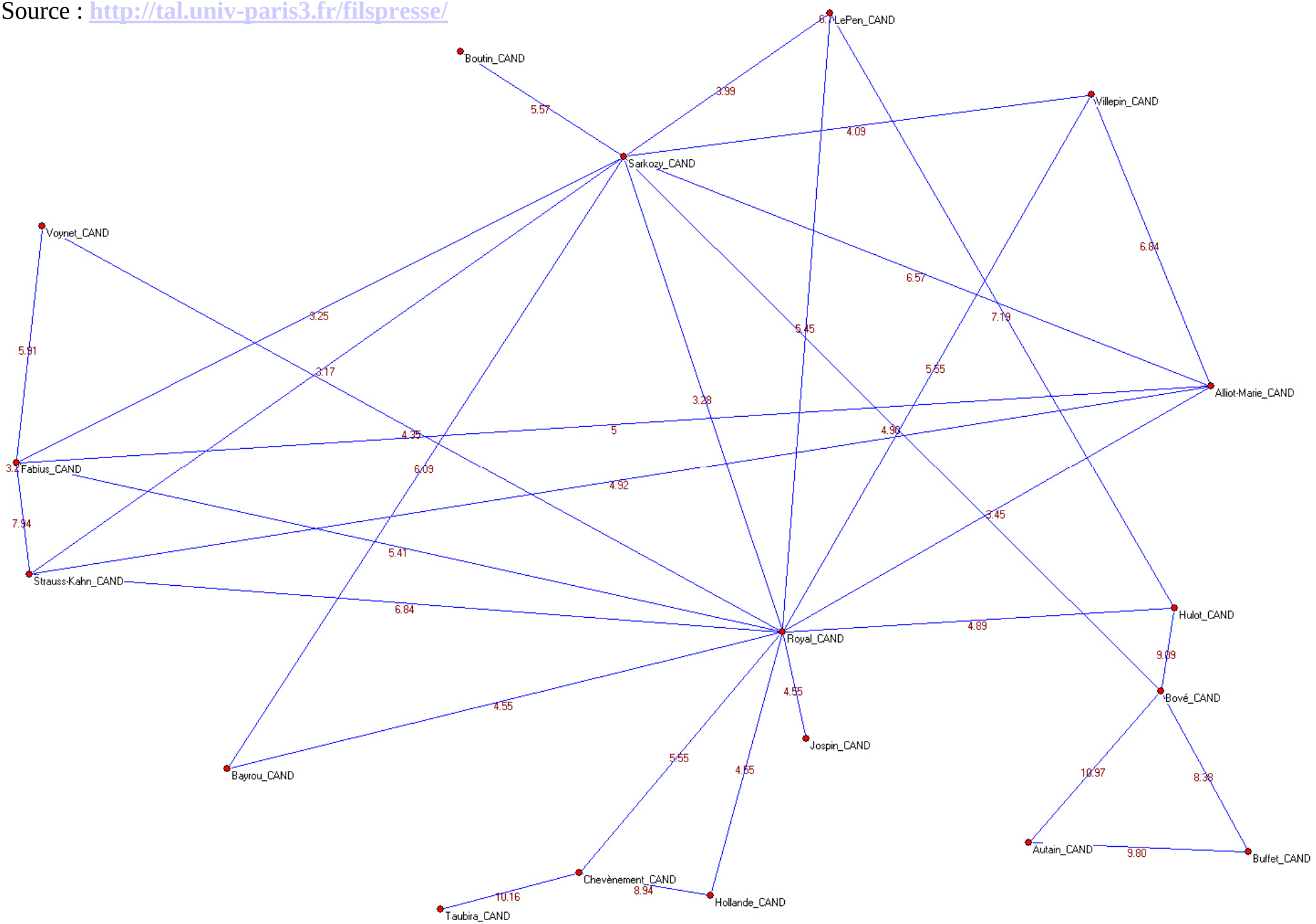
Les 17 fils RSS du journal *Le Monde* sur la période du 20/11/2006 au 21/12/2006.
Ces 17 fils ont été archivés tous les jours à 19h sur cette période.

| Rubrique | Fils RSS | Fils format texte (Lexico3) |
|---|---|---|
| A la Une | 0,2-3208,1-0,0.xml | 0,2-3208,1-0,0.txt |
| International | 0,2-3210,1-0,0.xml | 0,2-3210,1-0,0.txt |
| Europe | 0,2-3214,1-0,0.xml | 0,2-3214,1-0,0.txt |
| France | 0,2-3224,1-0,0.xml | 0,2-3224,1-0,0.txt |
| Société | 0,2-3226,1-0,0.xml | 0,2-3226,1-0,0.txt |
| Environnement | 0,2-3228,1-0,0.xml | 0,2-3228,1-0,0.txt |
| Entreprises | 0,2-3234,1-0,0.xml | 0,2-3234,1-0,0.txt |
| Médias | 0,2-3236,1-0,0.xml | 0,2-3236,1-0,0.txt |
| Rendez-vous | 0,2-3238,1-0,0.xml | 0,2-3238,1-0,0.txt |
| Sports | 0,2-3242,1-0,0.xml | 0,2-3242,1-0,0.txt |
| Sciences | 0,2-3244,1-0,0.xml | 0,2-3244,1-0,0.txt |
| Culture | 0,2-3246,1-0,0.xml | 0,2-3246,1-0,0.txt |
| Technologies | 0,2-651865,1-0,0.xml | 0,2-651865,1-0,0.txt |
| Cinéma | 0,2-3476,1-0,0.xml | 0,2-3476,1-0,0.txt |
| Voyages | 0,2-3546,1-0,0.xml | 0,2-3546,1-0,0.txt |
| Livres | 0,2-3260,1-0,0.xml | 0,2-3260,1-0,0.txt |
| Présidentielle 2007 | 0,57-0,64-823353,0.xml | 0,57-0,64-823353,0.txt |

SOURCE : http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html

Les candidats dans le fil ***Présidentielle 2007*** (période 18/10/2006-13/12/2006 – Catégorie : CAND)

Source : http://tal.univ-paris3.fr/filspresse/



# C’est quoi un fil RSS ?

- Illustration sur le *Monde.fr*
  - *Les fils RSS sont des flux de contenus gratuits en provenance de sites Internet. Ils incluent les titres des articles, des résumés et des liens vers les articles intégraux à consulter en ligne. Les dernières informations publiées sur Le Monde.fr peuvent ainsi venir enrichir automatiquement votre site Internet ou compléter vos sources d'informations déjà agrégées via un logiciel de lecture des flux RSS.*
- Liste des fils RSS du journal
  - *http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html*

# RSS : le tour du propriétaire

- Le standard RSS représente un moyen simple d'être tenu informé des nouveaux contenus d'un site web, sans avoir à le consulter.
- Le format « RSS » (traduisez « *Really Simple Syndication* ») permet ainsi de décrire de façon synthétique le contenu d'un site web, dans un fichier au format XML, afin de permettre son exploitation par des tiers. Le fichier RSS, appelé également flux RSS, canal RSS ou fil RSS, contenant les informations à diffuser, est maintenu à jour afin de constamment contenir les dernières informations à publier.
- Basiquement, un fil RSS est un fichier contenant le titre de l'information, une courte description et un lien vers une page décrivant plus en détail l'information. Cela permet à un site web de diffuser largement ses actualités tout en récupérant un grand nombre de visiteurs grâce au lien hypertexte permettant au lecteur de lire la suite de l'actualité en ligne.
- Les sites proposant un ou plusieurs fils d'actualités au format RSS arborent parfois un des logos suivants : 

# Un fil RSS

est un fichier XML répondant à quelques conditions simples de structure

# Structure

`<?xml version="1.0" ?>` déclaration de fichier xml

`<rss version="XX">` déclaration du fichier rss et de sa version

`....`

`</rss>` fin de fichier

# Structure

```
<?xml version="1.0" ?>
  <rss version="XX">

  <channel> déclaration du canal d'information
  Le channel (canal) permet de décrire le fil
  d'information de façon générale et permanente

  </channel>
   </rss>
```

# Un channel

```
<?xml version="1.0" ?>
  <rss version="XX">

  <channel>
<title>Le titre du fil</title>
<link>Le lien hypertexte général du fil (en général le
  site producteur ou un de ses chapitres)</link>
<description>Descriptif du fil</description>
<language>...</language>

...
  </channel>
   </rss>
```

# Les items

- Une fois qu'on a décrit le channel, apparaissent les items, les éléments documentaires essentiels qui vont composer le fil est qui sont le support des informations qui circuleront sur le fil.

# Les items

```
<?xml version="1.0" ?>
  <rss version="XX">
  <channel>
```

(description du channel)

```
      <item>
```
contenu de l'item (élément) n°1
```
      </item>
      <item>
```
contenu de l'item n°2
```
      </item>

....
      <item>
```
contenu du dernier item (en général n°10, parfois plus)
```
      </item>

</channel>
   </rss>
```

# Contenu des items

```
<?xml version="1.0" ?>

  <rss version="XX">
  <channel> (description du channel)
      <item>

         <title>Le titre de l'élément (de l'information) </title>

         <link>Le lien hypertexte menant directement à cette information
           </link>

         <description>Descriptif de l'information</description>

         <PubDate>date et heure</PubDate>

      </item>

      <item> (élément suivant)...


...
  </channel>
   </rss>
```

# Une application : un fil d'information presse avec un article

```
<?xml version="1.0" ?>

  <rss version="XX">
   <channel>
          <title>Mon journal préféré</title>
          <link>http://www.monjournal.fr/informations.html</lin
            k>
          <description>La meilleure source ...</description>
        <item>
          <title>La dernière info du jour</title>
          <link>Le lien hypertexte menant directement à cette information </link>
          <description>Descriptif de l'information</description>
          <PubDate>date et heure de mise en ligne de
            l'information</PubDate>
        </item>
        <item> (élément suivant: l'information précédente)...


...
   </channel>
    </rss>
```

# Fils RSS « A la Une » sur le Monde.fr

Channel

Item

Items

Zones textuelles traitées dans les BàO

```
<rss version="2.0">
 <channel>
   <title>Le Monde.fr : A la Une</title>
   <link>http://www.lemonde.fr</link>
   <description>Toute l'actualité au moment de la connexion</description>
   <copyright>Copyright Le Monde.fr</copyright>
  <image>
   <url>
     http://medias.lemonde.fr/mmpub/img/lgo/lemondefr_rss.gif
   </url>
   <title>Le Monde.fr</title>
   <link>http://www.lemonde.fr</link>
  </image>
  <pubDate>Fri, 01 Dec 2006 17:38:08 GMT</pubDate>
  <item>
   <title>Démonstration de force des prosyriens à Beyrouth</title>
   <link>
     http://www.lemonde.fr/web/article/0,1-0@2-3218,...
   </link>
   <description>
     Toutes les composantes de l&#39;opposition prosyrienne ont appelé à manifester dans le centre de Beyrouth pour réclamer la démission du premier ministre, Fouad Si...
   </description>
   <pubDate>Fri, 01 Dec 2006 16:05:25 GMT</pubDate>
   <guid isPermaLink="false">
     http://www.lemonde.fr/web/article/0,1-0@2-3218,36-840801,0.html?xtor=RSS-3208
   </guid>
   <enclosure url="http://medias.lemonde.fr/mmpub/edt/ill/2006/12/01/h_1_ill_841072_beyrouth.jpg" type="image/jpeg" length="2265"/>
  </item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
  + <item></item>
 </channel>
</rss>
<!--   -->
```

# Au travail !!!!
# BàO série 1…