

*Programmation et projet encadré*  
*Projet « Fil(s) de Presse »*  
*aka « Le projet Nuage »*

*(au départ des nuages de mots)*

# De la théorie...

## • Descriptif du cours

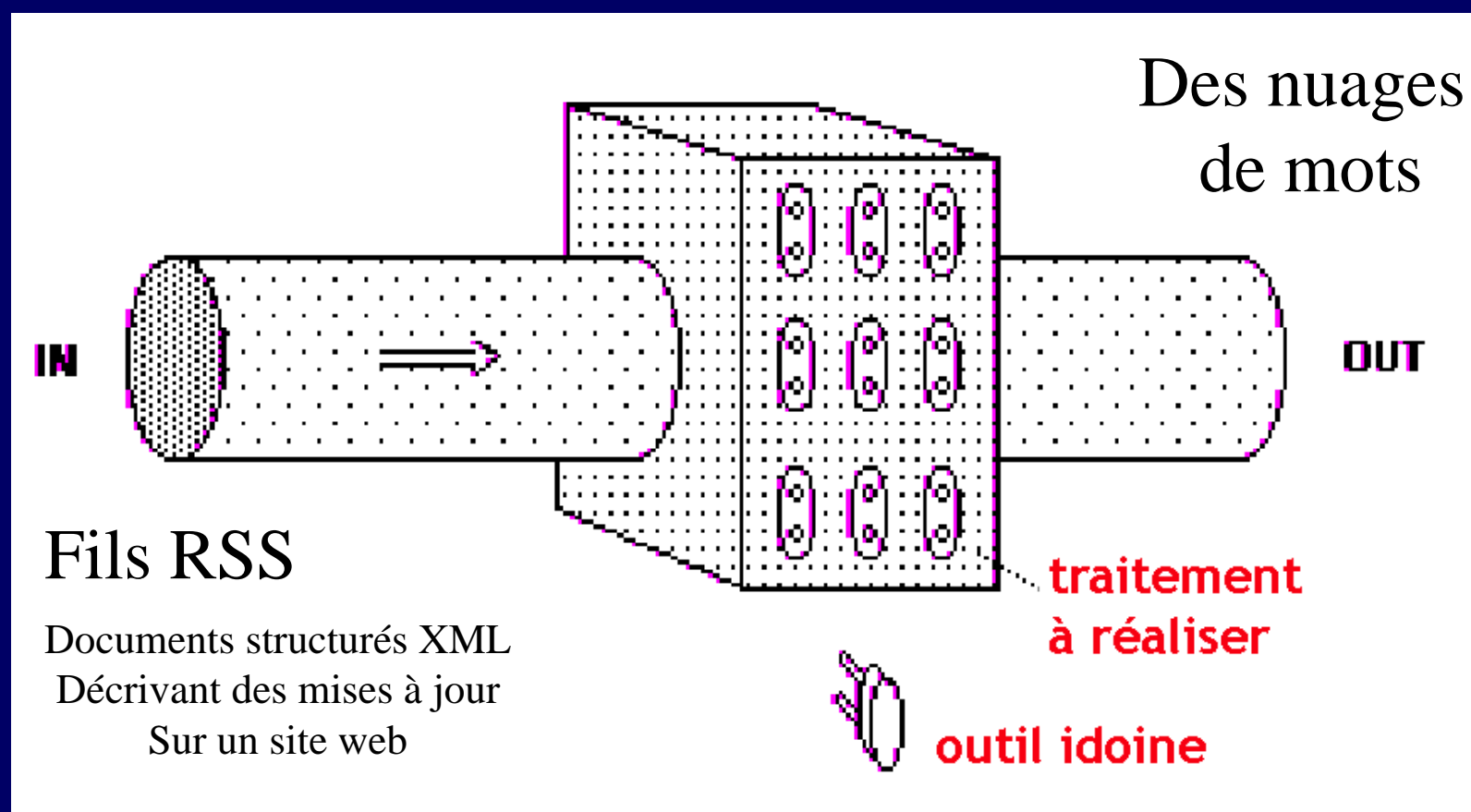
Mise en oeuvre d'une **chaîne de traitement textuel semi-automatique**, depuis la récupération des données jusqu'à leur présentation.

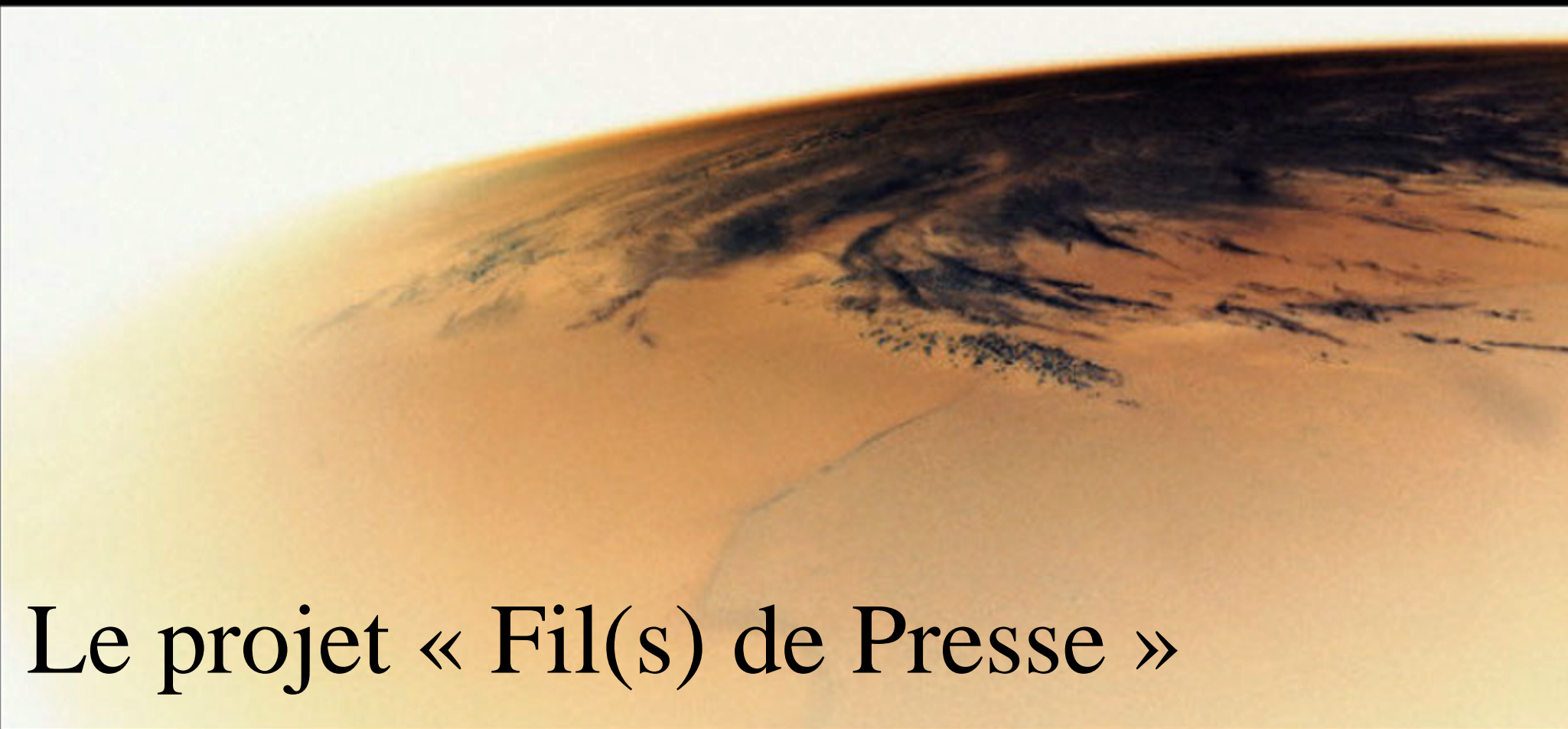
Ce cours posera d'abord la question des objectifs linguistiques à atteindre (lexicologie, recherche d'information, traduction...) et fera appel aux méthodes et outils informatiques nécessaires à leur réalisation (récupération de corpus, normalisation des textes, segmentation, étiquetage, extraction, structuration et présentation des résultats...).

Ce cours sera aussi l'occasion d'une évaluation critique des résultats obtenus, d'un point de vue quantitatif et qualitatif.

# A la pratique !

- « Cartographie » du projet encadré





# Le projet « Fil(s) de Presse »

- Parcours

- La vie des mots dans le Fils RSS :

- Présentation du projet ici :

<http://tal.univ-paris3.fr/filspresse/>

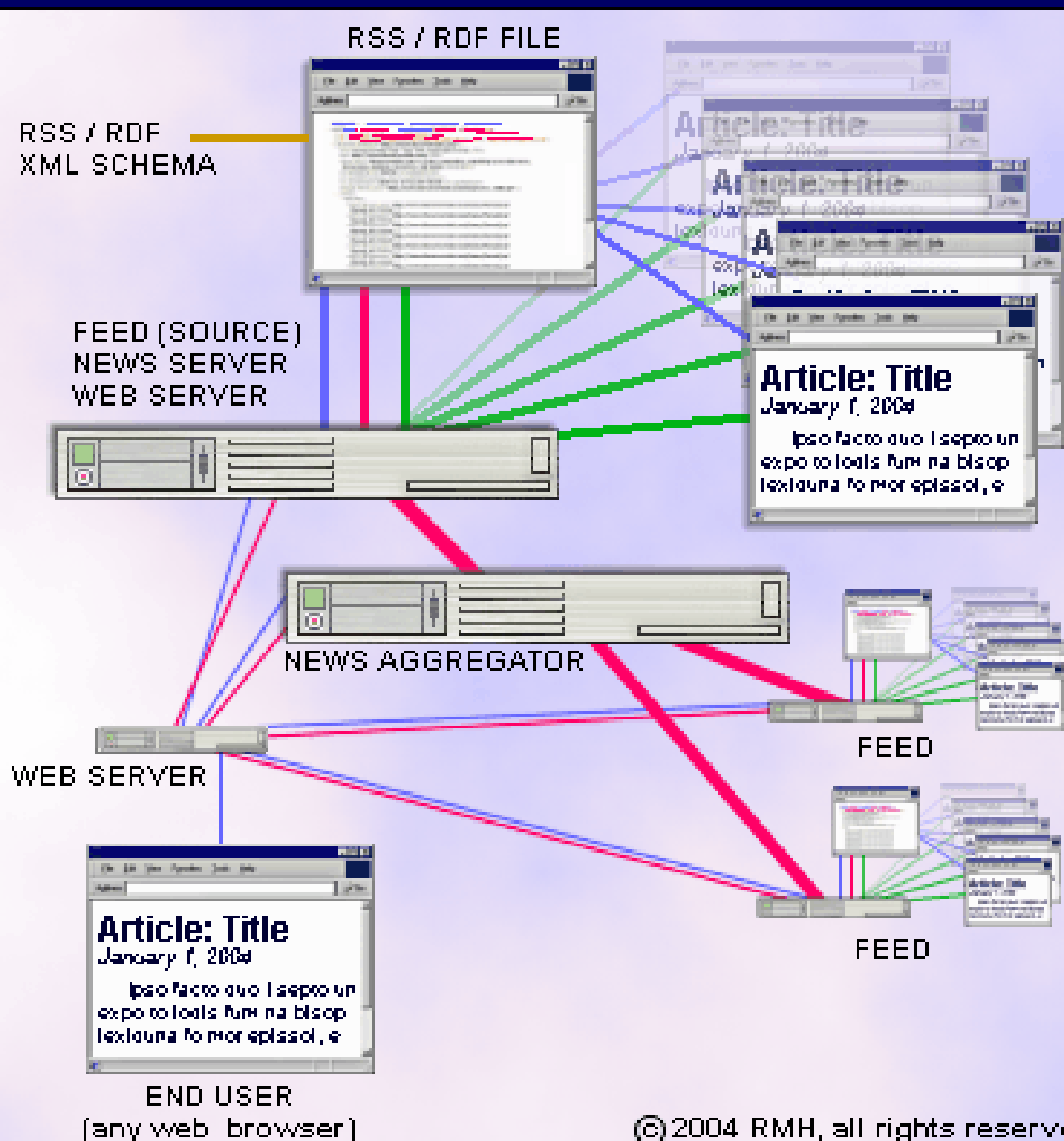
<http://tal.univ-paris3.fr/filspresse/projet-fils-de-presse.pdf>



# Introduction : RSS, qu'est-ce que c'est ?

- RSS est une norme mise au point en 1999 par les équipes de *Netscape* qui travaillaient sur des solutions permettant aux éditeurs de contenus de publier simplement leur travail sur Internet.
- Basé sur le langage XML, le **RSS** indexe en effet le contenu brut (sans s'occuper des données liées à sa forme) ce qui permet de le rendre malléable et de l'intégrer dans d'autres sites WEB par exemple. **RSS** signifiait ainsi "*Remote Site Syndication*". Netscape ayant abandonné son projet, les bloggers ont pris le relai et depuis **RSS** est plus souvent traduit par "*Really Simple Syndication*".
- 3 a 4 versions du format **RSS** circulent sur le WEB et l'enjeu est d'arriver à les standardiser. A quoi cela sert il ?
- Pour les webmasters le **RSS** est un moyen de récupérer du contenu d'autres sites pour animer son site web.
- Le format **RSS** présente un enjeu et un intérêt plus important pour l'internaute. En collant l'adresse d'un flux RSS dans son lecteur RSS, il reçoit directement les news du site dans son lecteur...c'est un peu le WEB qui vient a lui ("sorte de WEB push") avec par exemple le fil d'info de Libération <http://www.liberation.fr/rss.php>. Ce n'est pas un hasard si les journalistes et les veilleurs sont les premiers à l'avoir adopté.

# RSS



Comment ça marche ?

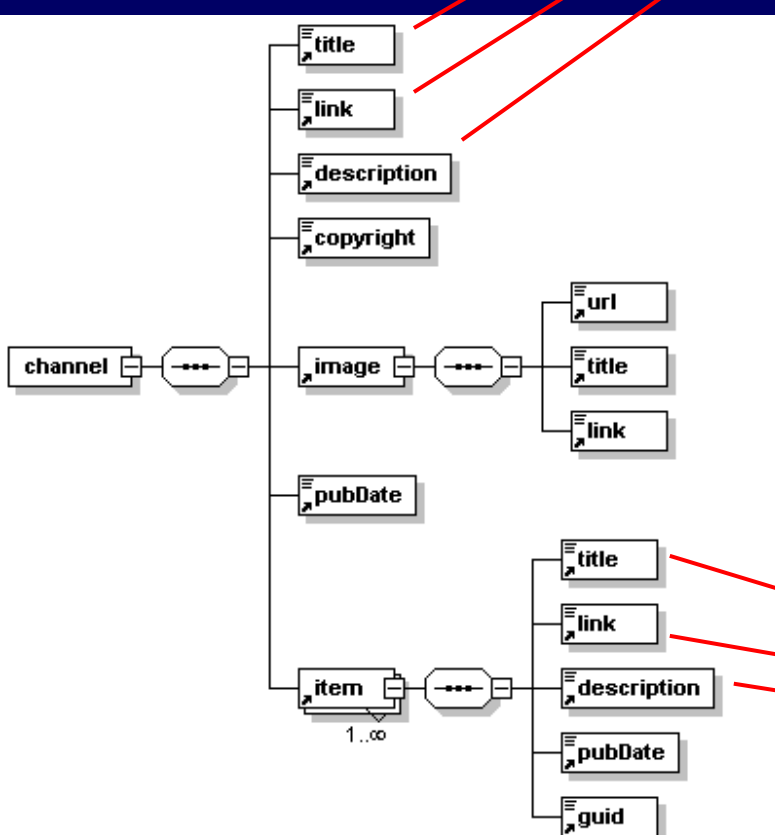
RSS est un format de diffusion (syndication) de contenus.

Le principe est simple : les sites/blogs mettent en place des flux RSS avec un format de données automatiquement structuré (en RDF ou en XML) et les utilisateurs peuvent les lire dans des outils dédiés (aggrégateurs, utilitaires mail, navigateurs).

# Un Fil RSS (un fil du Monde)

## Codage XML du FIL

### Arborescence du FIL

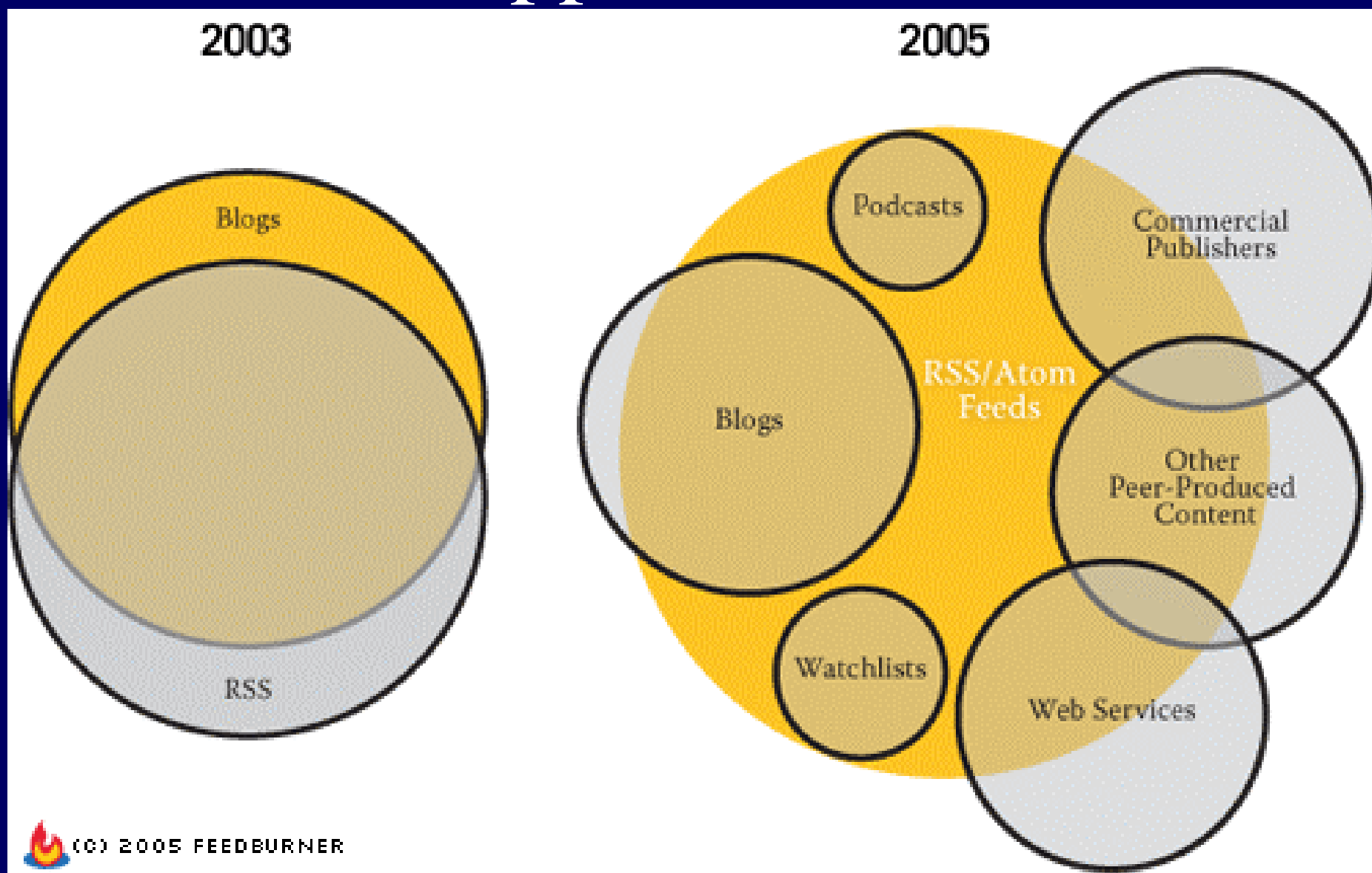


```

<?xml version="1.0" encoding="iso-8859-1"?>
<rss version="2.0" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <channel>
    <title>Le Monde.fr : International</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/ago/lemondefr_rss.gif</url>
      <title>Le Monde.fr</title>
      <link>http://www.lemonde.fr</link>
    </image>
    <pubDate>Tue, 22 Nov 2005 23:00:00 GMT</pubDate>
    <item>
      <title>Crispation du régime égyptien face aux succès électoraux des Frères musulmans</title>
      <link>http://www.lemonde.fr/web/article/0,1-0@2-3212,36-712875,0.html</link>
      <description>Les Frères musulmans ont consolidé le succès de la première phase des élections législatives, en décrochant treize nouveaux sièges.</description>
      <pubDate>Tue, 22 Nov 2005 14:45:03 GMT</pubDate>
      <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-0@2-3212,36-712875,0.html</guid>
    </item>
    <item>
      <title>Echange de tirs entre le Hezbollah et l'armée israélienne</title>
      <link>http://www.lemonde.fr/web/article/0,1-0@2-3218,36-712783,0.html</link>
      <description>Quatre combattants du Hezbollah libanais ont été tués au cours de violents affrontements avec l'armée israélienne dans le secteur controversé des Fermes de Chebaa.</description>
      <pubDate>Tue, 22 Nov 2005 07:24:18 GMT</pubDate>
      <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-0@2-3218,36-712783,0.html</guid>
    </item>
    <item>
      <title>Trois mois après son coup d'Etat, le président mauritanien plaide pour l'alternance politique</title>
      <link>http://www.lemonde.fr/web/article/0,1-0@2-3210,36-713134,0.html</link>
      <description>Dans un entretien au &#38;#34;Monde&#38;#34; et à Radio France internationale, le colonel Mohamed Vall confirme qu&#38;#39;il ne sera pas candidat à l&#38;#39;élection présidentielle de 2007.</description>
      <pubDate>Tue, 22 Nov 2005 16:07:57 GMT</pubDate>
      <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-0@2-3210,36-713134,0.html</guid>
    </item>
    <item>
      <title>Les Irakiens demandent un calendrier de retrait des Américains</title>
      <link>http://www.lemonde.fr/web/article/0,1-0@2-3218,36-713164,0.html</link>
      <description>La presse de Bagdad a salué, mardi 22 novembre, les résultats de la réunion inter-irakienne du Caire qui s&#38;#39;est achevée lundi, après trois jours de débats, par des résultats non négligeables.</description>
      <pubDate>Tue, 22 Nov 2005 14:51:26 GMT</pubDate>
      <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-0@2-3218,36-713164,0.html</guid>
    </item>
    <item>
      <title>Le président israélien accepte la dissolution de la Knesset</title>
      <link>http://www.lemonde.fr/web/article/0,1-0@2-3218,36-713164,0.html</link>
      <description>Le président israélien Moshé Katsav a annoncé qu&#38;#39;il acceptait la dissolution du Parlement, enclenchant le mécanisme qui conduira à des élections anticipées, vraisemblablement en mars.</description>
      <pubDate>Tue, 22 Nov 2005 18:36:31 GMT</pubDate>
      <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-0@2-3218,36-713164,0.html</guid>
    </item>
  </channel>
</rss>
  
```



# Développement de RSS



# Lectures

- **Wiki XWiki**

- <http://ceclines.xwiki.com/xwiki/bin/view/Main/Fils+RSS>
- Sur cette page de wiki, vous trouverez à peu près tout sur le RSS : La norme, comment utiliser les fils RSS, des articles, une sélection d'agrégateurs, mixer des fils RSS entre eux, mesurer l'activité des fils RSS, créer un fil RSS....

- **Fiche de l'ADBS consacrée au RSS**

- <http://www.adbs.fr/site/repertoires/outils/rss.php>

- **« RSS et la publication simultanée dans Internet », par Olivier Charbonneau**

- <http://www.culturelibre.ca/rss/>

- **« How feeds will change the way content is distributed, valued, and consumed », sur le weblog FeedBurner (Posted by Dick at November 21, 2005 10:47 AM )**

- <http://www.burningdoor.com/feedburner/archives/001518.html>

- **Sur le site URFIST Paris** (Unité Régionale de Formation à l'Information Scientifique et Technique) :

- « Modalités d'appropriation des fils RSS, blogs et wikis pour une veille informationnelle active » (<http://www.ccr.jussieu.fr/urfist/rss/>)

1. [Les fils RSS](#). [Les agrégateurs](#). [Les blogs](#). [Les wikis](#). [Les wikiblogs](#). [S'exercer à l'utilisation des fils RSS](#). [Ici](#) : avec les réponses

# Le projet « Fil(s) de presse »

## • Origine

- Le projet prend appui sur un programme qui est une implémentation en Perl d'une application présentée dans un tutorial rédigé par Jack Herrington sur le site d'IBM : "*Use PHP and XSL to create a DHTML link graph, Build an RSS parser that creates a keyword list with word frequencies*[1]", par Jack Herrington, Senior Software Engineer, Leverage Software, 4 octobre 2005.  
[1] <http://www-128.ibm.com/developerworks/edu/x-dw-x-lnkgrph-i.html?ca=drs-tp4005>

## • Objectif(s)

- Traitements de contenus textuels présents dans ces fils
- Réflexions et propositions de réalisations autour des traitements à mettre en œuvre pour cette application sur la base du prototype fourni (d'un point de vue traitement de flux)
- Réflexions et propositions de réalisations autour des problèmes de visualisations des informations dans ce type d'applications sur des données textuelles

# Le projet « Fil(s) de presse » (2)

- Votre point de départ
  - Un prototype d'archivage des fils
    - Le premier module du projet (« **Archivage des Fils de Presse** ») correspond au module permettant d'archiver les fils de manière continue et automatique afin de constituer la mémoire de ces fils.
  - Le programme « fil(s) de presse »
    - Le second module (« **Fil(s) de presse** ») correspond au module permettant de traiter un fil de presse donné (au format RSS) et de construire des traitements sur le contenu de ce fil (au départ, un nuage de mots).

# Archivage des fils

- Un processus expérimental a été mis en place pour archiver les fils de presse. L'idée est la suivante :
  - on a à disposition le corpus Le Monde depuis Avril 2003 (« *le Monde PROFOND* »)
  - on peut aussi avoir accès au fils RSS publiés quotidiennement (« *le Monde EN SURFACE* »)
- En archivant régulièrement les fils on a donc à portée de main le *PROFOND* et la *SURFACE*. Le processus mis en place aspire régulièrement les fils visés et crée des pages de navigation pour donner à voir les données archivées et les nuages de mots créés sur chacun des fils (*cf infra* le projet « Fil(s) de Presse » : programme construisant un nuage de mots à partir des contenus textuels présents dans un fil donné).
- Les données sont visibles provisoirement ici :
  - <http://sfmac.no-ip.com/fils-presse-arch/index.xml> (accès restreint)
- L'archivage mis en place concerne les fils du journal Le Monde et celui de l'AFP.

Fils

Pages  
navigation  
+ nuages

Nov	Aujourd'hui, 00:00	--	Dossier
19	samedi 19 n... 2005, 23:00	--	Dossier
20	dimanche 20... 2005, 23:00	--	Dossier
21	lundi 21 nov... 2005, 23:00	--	Dossier
22	mardi 22 no... 2005, 23:00	--	Dossier
23	mercredi 23 ... 2005, 23:00	--	Dossier
24	Hier, 23:00	--	Dossier
25	Aujourd'hui, 08:57	--	Dossier
00-00-00	Aujourd'hui, 00:01	--	Dossier
01-00-01	Aujourd'hui, 01:00	--	Dossier
02-00-00	Aujourd'hui, 02:00	--	Dossier
03-00-00	Aujourd'hui, 03:00	--	Dossier
04-00-00	Aujourd'hui, 04:00	--	Dossier
05-00-00	Aujourd'hui, 05:00	--	Dossier
06-00-00	Aujourd'hui, 06:00	--	Dossier
07-00-00	Aujourd'hui, 07:01	--	Dossier
08-00-00	Aujourd'hui, 08:00	--	Dossier
0,2-3208,1-0,0.xml	Aujourd'hui, 07:38	12 Ko	XML Pr...ist File
0,2-3210,1-0,0.xml	Aujourd'hui, 07:13	8 Ko	XML Pr...ist File
0,2-3214,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3224,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3226,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3228,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3234,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3236,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3238,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3242,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3244,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
0,2-3246,1-0,0.xml	Aujourd'hui, 07:13	4 Ko	XML Pr...ist File
08-00-00.html	Aujourd'hui, 08:00	8 Ko	HTML ...cument
AFP-stories.xml	Aujourd'hui, 07:52	8 Ko	XML Pr...ist File
fil1132902002-v1.xml	Aujourd'hui, 08:00	16 Ko	XML Pr...ist File
fil1132902002-v2.xml	Aujourd'hui, 08:00	24 Ko	XML Pr...ist File
fil1132902003-v1.xml	Aujourd'hui, 08:00	64 Ko	XML Pr...ist File
fil1132902003-v2.xml	Aujourd'hui, 08:00	108 Ko	XML Pr...ist File
nuage-afp-08-00-00.html	Aujourd'hui, 08:00	32 Ko	HTML ...cument
nuage-monde-08-00-00.html	Aujourd'hui, 08:00	132 Ko	HTML ...cument



# Le programme Fil(s) de Presse

- Le programme construit prend en entrée des fils RSS disponibles sur des sites de presse (Le Monde[1], Le Figaro[2], Libération[3]...) et produit des résultats donnant à voir :
  - - des **nuages de mots**
  - - une présentation des fils scrutés au format HTML et des comptages lexicométriques à partir des contenus textuels **des descriptions des articles** (disponibles dans les fils) mis à la disposition par les journaux.

[1] <http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html>

[2] <http://www.lefigaro.fr/xml/>

[3] <http://www.liberation.fr/page.php?Article=149907>

# Les nuages de mots

- Origine :
  - les TAGs de Technocrati
    - <http://www.technorati.com/tags/>
      - *Principe* :
        1. Arranging the words and terms in one paragraph, and
        2. Varying the font-sizes to represent the popularity of a keyword/ tag.
      - A tag is like a subject or category. This page shows the most popular 250 tags in alphabetical order. The bigger the text, the more active it is. [More Info »](#)  
<http://www.technorati.com/help/tags.html>
- Application : TagCloud
  - <http://tagcloud.com>
    - TagCloud is an automated [Folksonomy](#) tool. Essentially, TagCloud searches any number of RSS feeds you specify, extracts keywords from the content and lists them according to prevalence within the RSS feeds. Clicking on the tag's link will display a list of all the article abstracts associated with that keyword.
- Autres applications : *cf* texte du projet

Currently tracking 21.7 million sites and 1.7 billion links.

[Member Sign In](#)
[Sign Up](#)
[Help](#)
[Search](#)[Tags](#)[Blog Finder](#)[Popular](#)[About](#)[Options](#)

#### Sponsored Links

##### **Nobody reading your blog?**

Explode your Blog Traffic. 100% free blog traffic generator.  
[www.blogexplosion.com](http://www.blogexplosion.com)

##### **Blogging Evolved**

Elegant. Powerful. Professional. The better way to put a blog online  
[www.squarespace.com/](http://www.squarespace.com/)

##### **Car Crazy Community**

Upload your car videos and pics join groups, blogs, special events  
[www.CarCrazyCentral.com](http://www.CarCrazyCentral.com)

##### **Start your blog now**

Publish, be read, and get paid. Start writing instantly!  
[www.blogit.com/](http://www.blogit.com/)

##### **Ads by Google**

##### **Advertise on Technorati**

#### Most Popular

**News:** • CNN.com - 'Ugly dog' Sam dies at 14 - Nov...  
 • Guardian Unlimited | World Latest | Iraqi... • Cheney Accuses Iraq Critics of Shameless...

**Books:** • Harry Potter and the Goblet of Fire (Harry...  
 • Michael Langford's 35Mm Handbook • Mr. Benson

**Movies:** • Harry Potter and the Goblet of Fire... • Walk the Line (2005) • Pride & Prejudice (2005)

## Tags: The real-time web, organized by you

Currently tracking 3 million tags. Last updated 3:13 AM PST.

A tag is like a subject or category. This page shows the most popular 250 tags in alphabetical order. The bigger the text, the more active it is. [More Info »](#)

Show:

**A-Z**[All Languages](#)[About Me](#) ... [Actualité](#) ... [Actualités](#) ... [Actualités et politique](#) ...[Advertising](#) ... [Allmänt](#) ... [All Posts](#) ... [amazon????????](#) ... [Amigos](#) ... [amor](#) ...[Amusement](#) ... [Anime](#) ... [Announcements](#) ... [Apple](#) ... [Articles](#) ... [Asides](#) ...[Asterisk](#) ... [audio](#) ... [Babes](#) ... [Baby](#) ... [Baseball](#) ... [Blogs](#) ... [book](#) ...[books](#) ... [Bush](#) ... [Business](#) ... [Car](#) ... [Car Insurance](#) ... [Cars](#) ... [category](#) ...[Cell Phones](#) ... [China](#) ... [Cine](#) ... [cinema](#) ... [Comics](#) ... [Computadores e a](#) ...[Internet](#) ... [Computer](#) ... [Computers](#) ... [Computers and](#) ...[Internet](#) ... [Computing](#) ... [Cooking](#) ... [CSS](#) ... [Curiosidades](#) ... [Current](#) ...[events](#) ... [days](#) ... [Development](#) ... [diario](#) ... [Directory](#) ...[Divertissement](#) ... [Dogs](#) ... [dreams](#) ... [Entertainment](#) ...[Entretenimento](#) ... [Entretenimiento](#) ... [Environment](#) ...[ERROR: NOT PERMITTED METHOD: name](#) ... [etc](#) ... [Europe](#) ... [events](#) ...[EveryDay](#) ... [Everything](#) ... [F1](#) ... [fAcTs](#) ... [Family](#) ... [fashion](#) ... [Feeling](#) ...[Feelings](#) ... [FF11](#) ... [FFXI](#) ... [Film](#) ... [Films](#) ... [Firefox](#) ... [Flickr](#) ... [Food and](#) ...[Drink](#) ... [Football](#) ... [foreign-exchange](#) ... [Foreign Exchange](#) ... [Fotos](#) ...[Friends](#) ... [Fun](#) ... [Funny](#) ... [général](#) ... [Game](#) ... [Games](#) ... [Gaming](#) ...

## Formes : LE FIGARO

Tue Oct 25 19:59:02 2005

de la les à l le et d des du en un a une est

qui par sur dans pour Le au que son sont plus pas L se La Les  
aux s qu hier ce ne ont avec n été lui ses sa étudiants Il aujourd il après  
ministre tout ou mais deux être premier président ans leur pays emploi Pour A on  
depuis ils l 000 Un En gt comme tous lt septembre France Français Etat français  
sans européen où y faire nous Villepin très nouvelle leurs aussi européenne même ces  
sciences Jacques Une Europe elle déjà Chirac trois On encore soir non Dans cette 4  
entre 2 jour Ce dont toujours cinq bien Ils française 2004 autres C personnes millions  
Dominique ni 5 année occasion ancien projet national Paris grand 9 rien nombreux lui  
nombreuses Mais étudiant doit jours Nicolas monde selon rapport mode Sarkozy c dernier  
général François contre bac experts Union universitaires syndicats donc secrétaire octobre  
20 avait Constitution aviaire PS grippe nouveau aura coup étrangères moins tête visite  
technologies UMP succès désormais toute fait gouvernement chômage était peu pris loi  
Après meilleures place élèves qu'il Si ensemble faudra chez campagne République nombre  
négociations jusqu hommes b annoncé vie leader soit seul car environnement 25 bonne  
lundi ainsi Master palestinien chances oeuvre crise Asie Lyon bacheliers droit Deux voilà  
sera répondre mal 200 27 Figaro mardi nouveaux Elle mois autre fois jouer groupe  
régime acteurs 3 lors si cursus déclaré ouest Entre sondage compte inscriptions province  
première nom formation peut UDF avoir Irak faut dès 10 américain européens  
américains télévision certains notamment internationale devrait offre Française exportations  
pharmaceutique jeudi faits système chinois traité quatre espagnol Avec terre mobilisation  
audit socialiste B toutes défense Sécu stage font entreprises commerce cet thème filières  
candidats Au 30 député 2003 réalisé pire demande bord rapports 8 Syrie notre personne  
jeunes produits 700 études plusieurs unies économique solution loin majorité neuf Et trop  
temps 2006 trouver irakien famille devant maladie contrôle chacun référendum pendant dix

## nuage de mots sans lien

Dans cette première figure, le nuage de mots donne à voir l'ensemble des mots présents dans les descriptions des articles des fils d'un journal en ligne à un moment donné (ici Le Figaro).

## Formes

## Articles (titre, description et lien)

Tue Oct 25 19:59:02 2005

de la les à le et des du en un a  
 une est qui par sur dans pour Le au que son sont plus pas  
 L se La Les aux s qu hier ce ne ont avec n été lui ses sa étudiants il  
 aujourd il après ministre tout ou mais deux être premier président ans leur  
 pays emploi Pour A on depuis ils 1 000 Un En gt comme tous lt septembre  
 France Français Etat français sans européen où y faire nous Villepin très  
 nouvelle leurs aussi européenne même ces sciences Jacques Une Europe elle  
 déjà Chirac trois On encore soir non Dans cette 4 entre 2 jour Ce dont  
 toujours cinq bien Ils française 2004 autres C personnes millions Dominique ni  
 5 année occasion ancien projet national Paris grand 9 rien nombreux lui  
 nombreuses Mais étudiant doit jours Nicolas monde selon rapport mode  
 Sarkozy c dernier général François contre bac experts Union universitaires  
 syndicats donc secrétaire octobre 20 avait Constitution aviaire PS grippe  
 nouveau aura coup étrangères moins tête visite technologies UMP succès  
 désormais toute fait gouvernement chômage était peu pris loi Après meilleures  
 place élèves quil Si ensemble faudra chez campagne République nombre  
 négociations jusqu hommes b annoncé vie leader soit seul car environnement  
 25 bonne lundi ainsi Master palestinien chances oeuvre crise Asie Lyon  
 bacheliers droit Deux voilà sera répondre mal 200 27 Figaro mardi nouveaux  
 Elle mois autre fois jouer groupe régime acteurs 3 lors si cursus déclaré ouest

## 1 : Des balles contre des bulletins de vote

L'Irak actuel ne connaît ni la paix, ni la prospérité. La Constitution qui doit être votée aujourd'hui inclut cette règle des 25%, mais n'a rien de libéral...

(<http://www.lefigaro.fr/debats/20051015.FIG0154.html>)

## 2 : La ratification devrait continuer, malgré les deux «non»

Encore sous le choc du non français, les dirigeants européens ont enregistré une nouvelle déception avec le rejet massif de la Constitution aux Pays-Bas hier, mais la plupart d'entre eux veulent s'accrocher à la poursuite du processus de ratification du traité...

(<http://www.lefigaro.fr/referendum/20050603.FIG0132.html>)

## 3 : Gérard Chaliand : «Une Constitution irakienne au forceps»

LE FIGARO. - Les Irakiens sont consultés aujourd'hui par référendum sur une Constitution. Comment évaluez-vous les risques politiques et sur le terrain ? GÉRARD CHALIAND...

(<http://www.lefigaro.fr/debats/20051017.FIG0356.html>)

## 4 : Irak : l'autre réalité

Dur à avaler pour les anti-Bush : les Irakiens acceptent la démocratie offerte par les Etats-Unis. Samedi, ils ont été 61% à participer au référendum sur la Constitution. Le 30 janvier, ils s'étaient pareillement mobilisés, malgré les menaces, pour être leurs députés...

(<http://www.lefigaro.fr/debats/20051021.FIG0193.html>)

## 5 : Redéfinir les termes du débat sur l'Europe...

Le rejet de la Constitution pour l'Europe par les électeurs français et néerlandais a indéniablement causé un sérieux retard au processus d'intégration européenne...

(<http://www.lefigaro.fr/debats/20051019.FIG0138.html>)

## nuage de mots avec liens

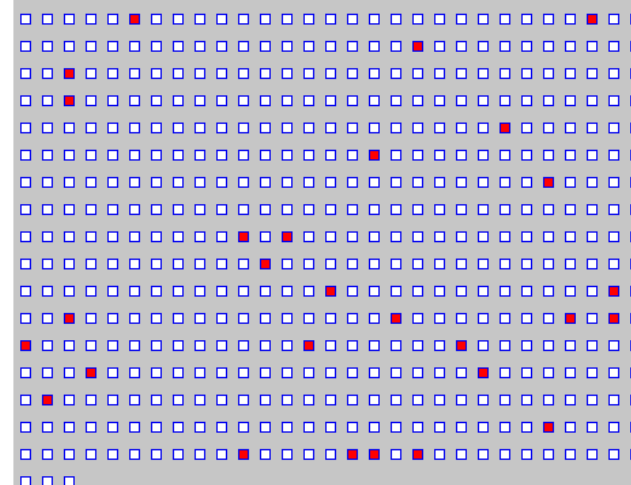
Dans la seconde, on peut voir un nuage similaire dans lequel chaque mot donne accès *via* un clic aux contextes dans lesquels ce mot apparaît (colonne de droite) : le contexte est constitué par le titre de l'article, sa description et son URL.

## Formes

UN CARRE = UN ARTICLE [■ = le mot est dans l'article ; □ = le mot n'y est pas.]

Sat Nov 5 17:26:47 2005

de la et le à les l d des un en du une  
est qui a pour dans sur au plus que par se Le La s il son pas ou avec Les sont ce  
sa aux ont L ses ne ans qu deux tout n mais étudiants fait leur il A comme cette été vous lui y 000 Pour on premier  
encore ces Une C sans En aussi déjà gt leurs lt Paris année président être hier bien Un après faire France aujourd entre ils  
nouveau c elle tous même jeunes français depuis dont Mais quatre dernier chez vie directeur trois nous autres où souvent ministre très peu lui  
mode quelques Dans prix sera cinq université monde était Au Cette avant Jacques groupe si va On sciences Europe Après 1 Et temps Marseille  
emploi fin mois avoir entreprises octobre bon droit moins autre gouvernement jusqu tête toujours grand 2005 Ils genre universités euros grands  
projet dix novembre De non formation petit loi place 5 Français politique désormais amp selon point près travail Ce écoles dès contre bac Si  
recherche celle européen toute première pays 2004 font Dominique 8 hommes 2 crise ville jour grandes 3 tendance autant femmes peut création  
nouvelle vient Depuis sous homme Villepin années notamment trop euro Figaro société votre prochain origine exposition nouvelles soit car  
enseignement D nombreux Entre six public Aujourd marché partie 20 réunion devant environ possible mettent Etat reste mise étudiant française  
américain rapport Toulouse enquête centrale eux scène Avec Etats-Unis notre Plus hausse M celui belle nombreuses fête jours cours ici grande ailleurs  
cet supérieur septembre villes syndicats p avait b e mal nouveaux an 9 jeune vers également soir compte partir voir doit réforme cas livre question  
logement programme - européenne Pierre 4 jamais nombre annoncé 2006 chacun auteur expérience Elle haut rien diplômés fois déficit direction choix  
raison célèbre nos Pas chercheurs général stage spécialisé effet derniers plusieurs 26 marque presque pourtant parfois br contrat 27 succès mettre lors  
part déclaré film saison parisienne offre Sarkozy chef histoire sociales passe personnes Alors François siècle gestion universitaire sortie 30 surtout bord  
situation universitaires vos 25 quotidien pendant rentrée heure Des je national voix mardi donner jouer seraient femme côté musique heures début Internet  
tu système financement conseil devient millions Du issue experts études signe loin met coût trouver enfant monte autour Master action forme Brême 21



Dans la troisième, on y voit toujours le même nuage de mots sur la gauche, dans lequel chaque mot donne accès *via* un clic à une « *représentation cartographique*[1] » du contenu du fil scruté dans laquelle le contenu textuel de la description d'un article est représenté par un carré, les articles contenant le mot cliqué sont associées à des carrés rouges et les autres à des carrés blancs . Chaque carré est donc associé à un article en ligne (si on clique sur le carré on accède à l'article en ligne).

[1] Ce développement s'inscrit dans les travaux faits autour de Lexico3 pour construire des représentations des textes donnant à voir les unités textuelles manipulées à travers des objets graphiques :

<http://lexico3.no-ip.org/>, <http://tal.univ-paris3.fr/CE-query/>



# Développements à construire

- Traitements des contenus des fils
  - Etiquetage, repérage de syntagme, segments répétés etc.
- Réflexion sur le type de sorties à construire avec ce type de données
- Archivages des fils de presse (BDD, XML...)

*(après prise en main de l'ensemble de la chaîne...)*

# Avant de partir, je prépare un journal de bord

- Votre parcours doit être effectué en établissant un « log-book » retraçant l'ensemble de vos activités
  - Blog (*cf* Blog (*pluri*)TAL)
    - Si vous faites ce choix, vous disposerez d'un compte sur le weblog pour le projet mené
    - Vous pourrez ensuite publier des « brouillons » ou des billets officiels retraçant vos activités
  - Etc.

# un journal de bord sur le blog (pluri)TAL (1/2)

http://tal.univ-paris3.fr/blogtal/

(pluri)TAL



Journal de lectures, de liens, d'activités pour les étudiants du secteur TAL [Université Paris 3 Sorbonne nouvelle | ILPGA] HyperToile : <http://tal.univ-paris3.fr>

11/10/2005

Human Rights Corpus / Corpus Droits de l'Homme, v.1 (XML-TEI)

Rubrique(s) : [Bookmark](#) [XML](#) [TAL](#) [Ressources](#) [WWW](#) [Corpus](#)  
Auteur : SFA | Heure : 3:33 pm

Message diffusé par la liste [Langage Naturel LN@cines.fr](mailto:Langage.Naturel@lncines.fr) :

We are happy to announce the release of the 'Human Rights Corpus / Corpus Droits de l'Homme, v.1, available on our web site : Université de Paris 13 - Laboratoire de Linguistique Informatique

<http://www-lli.univ-paris13.fr/ressources>

The corpus is composed of 28 International Conventions, from 1948 (Universal Declaration of Human Rights) up to 2000. The choice of the texts has been made with an expert of the field, with the aim to have a representative view of the Human Rights reference texts and of the language and vocabulary used.

Each text is given in 2 or 3 languages : English and French, and Spanish when the Convention is one of the United Nations. These versions are aligned at the level of the finest subdivision (article) through an appropriate design of identifiers. **The encoding is in XML and follows the guidelines of the TEI.** A special attention has been devoted to the realization of the Header ; in particular, the "TagUsage" part is fully developed in order to make understandable the choices made for the encoding and the meaning of each XML/TEI tag in our context. Please contact us to let us know your interests or remarks : [corpus@lli.univ-paris13.fr](mailto:corpus@lli.univ-paris13.fr).

Fabrice ISSAC, Computational Linguist, Christine CHODKIEWICZ, Lawyer and Linguist, Bénédicte PINCEMIN, Linguist

Comments Off

Enquête sur l'utilisation des logiciels libres dans les collectivités territoriales

Rubrique(s) : [Bookmark](#) [Outils](#) [Lectures](#) [Informatique](#)  
Auteur : SFA | Heure : 8:31 am

Sur la site @netville : Le principe de cette enquête, réalisée par la Mission Ecoter et l'Apronet, est basé sur un questionnaire en ligne et porte sur l'utilisation (et non l'usage) de logiciels libres dans les collectivités. Le [rapport au format PDF](#) : Editeur(s) : Mission ECOTER/APRONET. Année : 2005. Document : 34 pages - PDF - 505 Ko. Disponibilité : en téléchargement

Comments Off

Contact : Serge Fleury

[serge.fleury@univ-paris3.fr](mailto:serge.fleury@univ-paris3.fr)  
[sfweb.no-ip.org](http://sfweb.no-ip.org)

ATONET

wiki (TAL-Lexicométrie)

Recherche :

Search

Rubriques (pluri)TAL

[General](#)  
[Emplois](#)  
[Bookmark](#)  
[Conférences](#)  
[Lectures](#)  
[Référence Bibliographique](#)  
[Linguistique-lectures](#)  
[Informatique-lectures](#)  
[BLOGs](#)  
[Web Sémantique](#)  
[XML](#)  
[RDF](#)  
[XSL](#)  
[OWL](#)  
[Métadonnées](#)  
[RSS](#)  
[TAL](#)  
[Outils-TAL](#)  
[Lectures-TAL](#)  
[Ressources](#)

Des billets réguliers  
Classés par rubrique

Des archives  
(plus bas sur la page)

[sommaire](#)

# un journal de bord sur le blog (*pluri*)TAL (2/2)

Une plateforme d'édition en ligne

The screenshot shows the WordPress 'Write' page. The top navigation bar includes links like 'Write', 'Edit', 'Categories', 'Links', 'Users', 'Backup/Restore', 'Options', 'Plugins', 'Templates', 'Profile', 'Wpstats', 'View site', and 'Logout'. The 'Logout' link is circled in red. Below the navigation bar, the 'Title' field is labeled 'Titre du billet'. The 'Post' content area is labeled 'On rédige le billet ici...'. Above the content area, the 'Quicktags' section includes buttons for 'str', 'em', 'link', 'b-quote', 'del', 'ins', 'img', 'ul', 'ol', 'li', 'code', 'more', 'page', 'Dict.', and 'Close Tags'. This section is labeled 'Quelques outils d'édition...'. To the right, the 'Categories' sidebar lists various categories like 'Audio', 'MUSIC', 'BLOGs', 'Bookmark', 'Conférences', 'E-Learning', 'Emilie's Work', 'Emplois', 'Enseignement', 'Cours', '2004-2005', 'Cours', '2005-2006', 'Cours en ligne', 'Cours M1', 'Cours M2', 'Cours TAL', 'P3', and 'Cours-L1'. This sidebar is labeled 'On classe le billet...'. At the bottom, the 'Publish' button is circled in red, and the 'Advanced Editing »' link is also circled in red. This area is labeled 'On publie ici : en ligne ou un brouillon'. The 'Advanced Editing' link is further labeled 'Paramétrage avancé pour l'édition : Prévisualisation, Acceptation des commentaires...'. The 'TrackBack' section at the bottom is labeled 'TrackBack an URL: (Separate multiple URIs with spaces.)'.

WordPress

You're lookin' swell, Dolly

Write Edit Categories Links Users Backup/Restore Options Plugins Templates Profile Wpstats View site Logout

Connexion sur un compte prédéfini

Titre du billet

Post

Quicktags: **str** em link b-quote del ins img ul ol li code more page Dict. Close Tags

On rédige le billet ici...

Quelques outils d'édition...

On classe le billet...

On publie ici : en ligne ou un brouillon

Paramétrage avancé pour l'édition : Prévisualisation, Acceptation des commentaires...

Categories

- ☐ Audio
- ☐ MUSIC
- ☐ BLOGs
- ☐ Bookmark
- ☐ Conférences
- ☐ E-Learning
- ☐ Emilie's Work
- ☐ Emplois
- ☐ Enseignement
- ☐ Cours
- 2004-2005
- ☐ Cours
- 2005-2006
- ☐ Cours en ligne
- ☐ Cours M1
- ☐ Cours M2
- ☐ Cours TAL
- P3
- ☐ Cours-L1

☒ PingBack the URIs in this post ?

TrackBack an URL: (Separate multiple URIs with spaces.)

Save as Draft Save as Private Publish Advanced Editing »

# Le weblog du projet « Fil(s) de Presse »

- <http://tal-p3.wordpress.com>
  - Login : ...
  - Mot de passe : ...

*pluriTAL*

October 21, 2005

## Démarrage...

Filed under: [Uncategorized](#) — tal-p3 @ 6:57 pm

Ouverture prochaine...



Comments Off

Powered by [WordPress](#)

### blogroll

[wordpress.com](#)  
[wordpress.org](#)

### master tal recherche

[blog \(pluri\)tal](#)  
[site plurital](#)

### categories:

[uncategorized](#)

### search:

### archives:

[october 2005](#)

### meta:

[login](#)  
[rss](#)  
[comments rss](#)  
[valid xhtml](#)  
[xfn](#)  
[wp](#)

- Bon travail....