



L7T04/L8T07
(Master TAL 1/2)
Paris 3 - ILPGA / Paris X / INALCO

Serge Fleury
serge.fleury@univ-paris3.fr

Sommaire

1	Table des figures	3
2	Descriptif de cours	4
3	Projet : nuage de mots dans les fils de news sur un corpus de presse	5
3.1	Préambule	5
3.2	Composantes du projet	5
3.3	La plateforme d'archivage des Fils de Presse	6
3.4	Le projet Fil(s) de Presse	7
3.5	Architecture du projet « nuage de mots »	9
3.6	Fils de Presse	13
3.6.1	Le Monde	13
3.6.2	Libération	14
3.6.3	Le Figaro	15
4	Liens/projets	16
4.1	Les nuages de Tags chez Technocrati	16
4.2	Annuaire de Fils	17
4.3	AlertInfo, un agrégateur RSS de la presse française	17
4.4	Amazon concordance	19
4.5	TagClouds (« Nuage de mots »)	22
4.6	Le filtre Google	24
4.7	10x10 : images du monde	26
4.8	Projet « Post Remix » (<i>Washington Post</i>)	28
4.9	1000Tags	30
4.10	ZoomTags	31
5	Lectures	32
5.1	Conversation : De la représentation visuelle à la complexité documentaire	32
5.2	Blog <i>Technologies du Langage</i> (par Jean Véronis)	32
5.3	Bibliothèque 2.0	33
6	Les développements à construire	35
6.1	Traitements des contenus des fils	35
6.2	Archivages des fils de presse (<i>in-progress</i>)	35
6.3	En Vrac	35
6.3.1	Commentaires AS	35
7	Liens et développements autour de RSS	37

1 Table des figures

Figure 1 : Schéma "Le projet Nuage"	4
Figure 2 : Archivage des fils, arborescence	6
Figure 3 : Nuage de mots sans lien	7
Figure 4 : Nuage de mots avec liens	8
Figure 5 : Nuages de mots avec "carte des sections" (1 section = 1 carré = 1 article).....	8
Figure 6 : Architecte initiale "en amont"	9
Figure 7 : Architecture modifiée "en amont"	10
Figure 8 : Schéma du lexique construit.....	10
Figure 9 : Architecture "en aval"	11
Figure 10 : Les fils du Monde	13
Figure 11 : Le Fil de Libé.....	14
Figure 12 : Les Fils du Figaro	15
Figure 13 : Nuages de TAG	16
Figure 14 : Amazon "In the Beginning...was the Command Line"	19
Figure 15 : Menu "concordance" sur Amazon	20
Figure 16 : Nuage de mots "concordance"	20
Figure 17 : Contextes "Concordance"	21
Figure 18 : projet TagCloud	22
Figure 19 : tagcloud sur Fils du Monde	22
Figure 20 : Paramétrage du tagcloud LeMonde	23
Figure 21 : Projet NewsMap	24
Figure 22 : Projet NewsMap (france).....	25
Figure 23 : Projet 10x10.....	27
Figure 24 : Projet NewsCloud (Washington Post).....	29
Figure 25 : des Tags pour visualiser des collections de bibliothèques.....	34

2 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.

<http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/masterproj.htm>

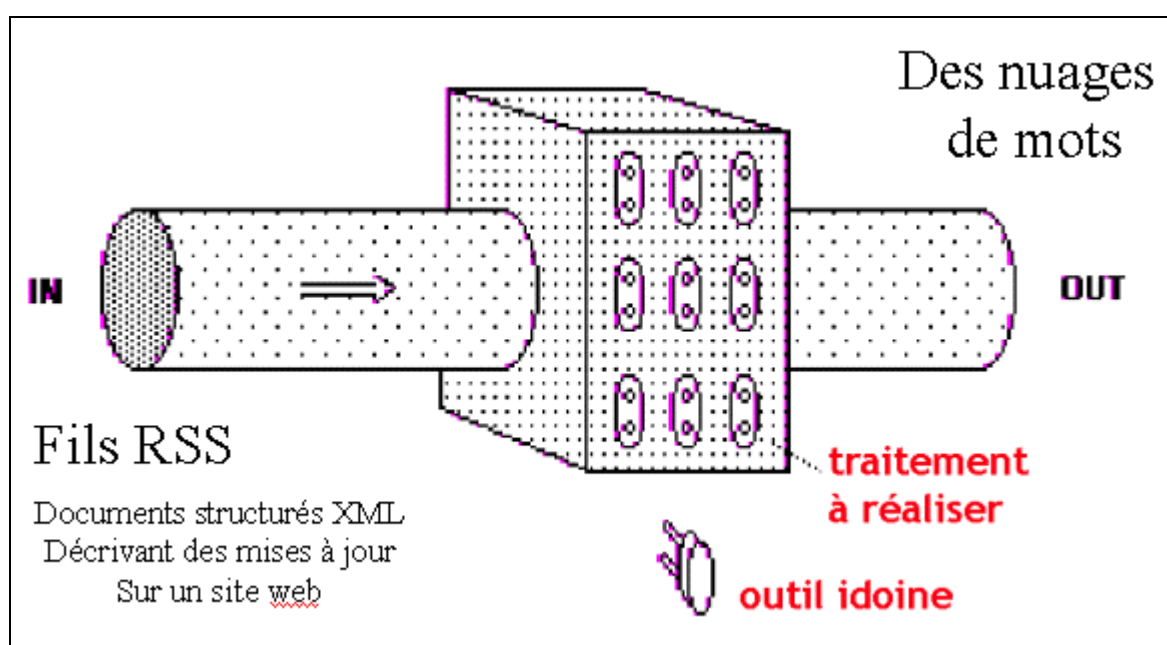


Figure 1 : Schéma "Le projet Nuage"

3 Projet : nuage de mots dans les fils de news sur un corpus de presse

Le projet en cours de développement est visible à cette adresse :

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

Programme de lecture de fils RSS¹ d'organe de presse en ligne.

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.

3.1 Préambule

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 frequencies²", par Jack Herrington, Senior Software Engineer, Leverage Software, 4 octobre 2005. (désormais note [Herrington, 2005])

Abstract : *In this tutorial, you learn to build a link graph with XML, PHP, and JavaScript code. Link graphs are paragraphs of keywords in which the font size of each word is determined by some data value -- in this case, the frequency of the term. The more often the term occurs, the larger the font size of the word. This tutorial shows how to build an RSS parser that in turn builds a keyword list along with the word frequencies. It also demonstrates how to use XSLT to create an HTML page that shows the link graph and relates its term to its original article.*

3.2 Composantes du projet

Le projet est composé de 2 modules.

Le premier (« **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).

Le second (« **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.

¹ 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 (agrégateurs, utilitaires mail, navigateurs). Cf cours URFIST <http://www.ccr.jussieu.fr/urfist/rss/>

² <http://www-128.ibm.com/developerworks/edu/x-dw-x-lnkgrph-i.html?ca=drs-tp4005>

3.3 La plateforme d'archivage des Fils de Presse

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 2003³ (« *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 (*cf infra*) et celui de l'AFP⁴. La figure suivante donne une représentation de l'organisation de cet archivage :

Arborescence	Contenu	Taille	Type
Nov	Aujourd'hui, 00:00	--	Dossier
19	samedi 19 nov... 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

Figure 2 : Archivage des fils, arborescence

Le processus d'archivage est déclenché toutes les heures et produit à chaque lancement un archivage des fils, des pages de navigation et les données nécessaires pour construire les nuages de mots.

³ <http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm>

⁴ <http://www.afp.fr/francais/rss/stories.xml>

3.4 Le projet Fil(s) de Presse

Le programme construit prend en entrée des fils RSS disponibles sur des sites de presse (Le Monde⁵, Le Figaro⁶, Libération⁷...) 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.

Les figures suivantes présentent les différents types de nuages construits :



Figure 3 : 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).

⁵ <http://www.lemonde.fr/web/rss/0.48-0.1-0.0.html>

⁶ <http://www.lefigaro.fr/xml/>

⁷ <http://www.liberation.fr/page.php?Article=149907>



Figure 4 : 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.



Figure 5 : Nuages de mots avec "carte des sections" (1 section = 1 carré = 1 article)

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 cartographique⁸ » du contenu du fil scruté

⁸ 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/>

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).

Dans les trois figures, la taille de la police de caractères utilisée pour afficher le mot dans le nuage est déterminée par la fréquence du mot dans l'ensemble des articles scrutés pour un journal donné.

3.5 Architecture du projet « nuage de mots »

Dans le projet initial [Herrington, 2005], l'architecture « en amont » de l'application a l'allure suivante :

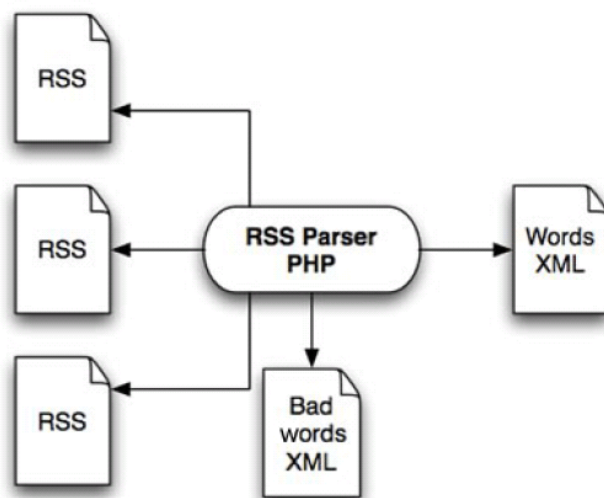


Figure 6 : Architecte initiale "en amont"

L'application lit des flux RSS et déclenche un *parser* RSS (écrit en PHP) qui a pour tâche de sélectionner les zones de texte à explorer puis de lancer une opération de segmentation de ces contenus textuels en ne retenant que les mots non présents dans une liste prédéterminée (mots vides).

L'architecture maintenue pour le projet présenté ici est la suivante :

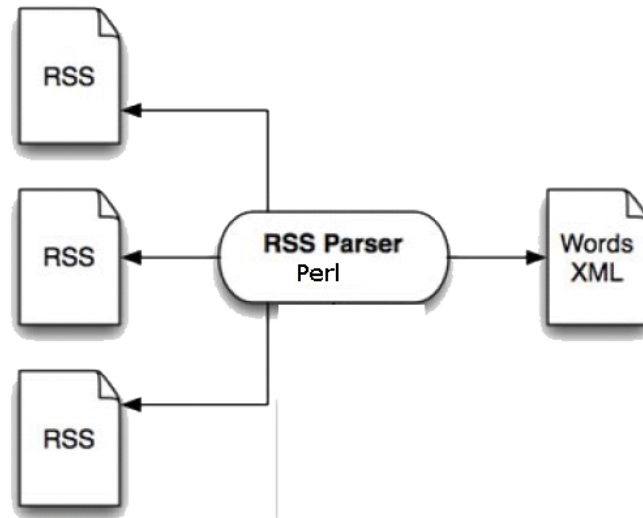


Figure 7 : Architecture modifiée "en amont"

Le principe général est conservé, tout le code est réécrit en Perl, *parser* compris. Tous les mots présents dans les contenus textuels scrutés sont conservés. Les mots retenus et comptés sont sauvegardés au format XML. Le fichier produit a l'allure suivante :

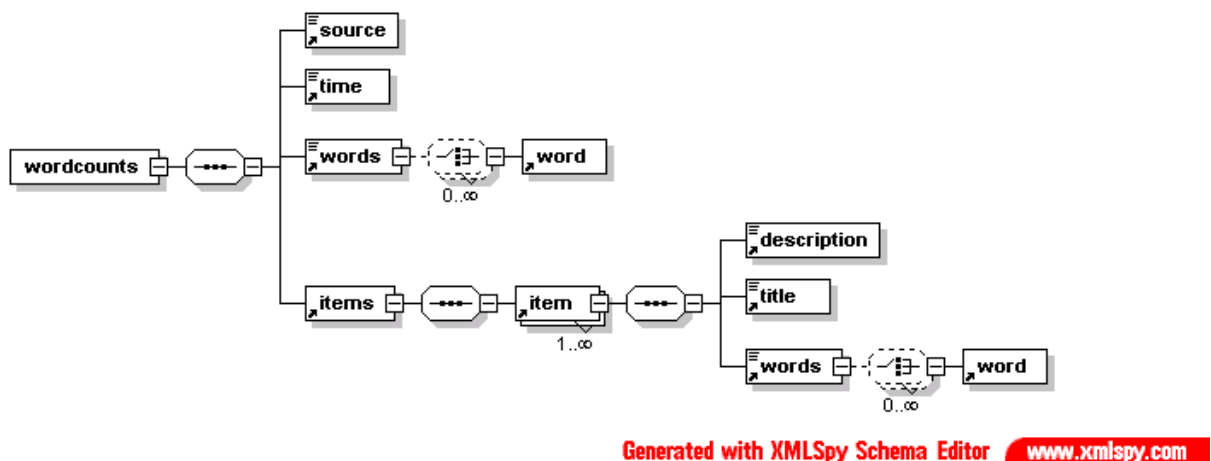


Figure 8 : Schéma du lexique construit

Dans ce schéma, l'élément `words` contient la liste de tous les mots (et leur fréquence) pour un fil de presse donné, l'élément `item` contenant la liste de tous les mots pour un article donné contenu dans ce fil.

On présente ci-dessous un extrait du lexique construit :

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet type="text/xsl" href="parsersss.xsl"?>
<wordcounts>
<source>LIBERATION</source>
<time>Wed Oct 26 08:06:28 2005</time>
<words>
<word text="de" count="16" />
<word text="le" count="5" />
<word text="la" count="5" />
<word text="d" count="5" />
...
</words>
<items>
```

```
<item url="http://www.liberation.fr/page.php?Article=332624" title="">
<description><![CDATA[Mort du dessinateur aux personnages diaphanes et angéliques, rendu
célèbre par un générique d'Antenne 2.]]></description>
<title><![CDATA[Feu Folon]]></title>
<words>
<word text="Mort" text2="Mort" />
<word text="du" text2="du" />
<word text="dessinateur" text2="dessinateur" />
<word text="aux" text2="aux" />
<word text="personnages" text2="personnages" />
<word text="diaphanes" text2="diaphanes" />
<word text="et" text2="et" />
<word text="angéliques" text2="angeliques" />
<word text="rendu" text2="rendu" />
<word text="célèbre" text2="celebre" />
<word text="par" text2="par" />
<word text="un" text2="un" />
<word text="générique" text2="generique" />
<word text="d" text2="d" />
<word text="Antenne" text2="Antenne" />
<word text="2" text2="2" />
</words>
</item>
...
</items>
</wordcounts>
```

Une modification mineure a été réalisée dans la grammaire du fichier lexique produit par rapport à l'application initiale. La présence de caractères accentués dans les mots posant des problèmes pour la seconde partie de l'application (celle utilisant le script établissant le lien entre le mot et ses contextes), un attribut a été ajouté dans les éléments décrivant les mots, celui-ci contenant après transcodage, la forme graphique normalisée du mot sans caractères accentués (générique est réécrit generique).

Dans un deuxième temps, l'application construit le nuage des mots en utilisant le lexique produit et en appliquant sur la sortie XML contenant ce lexique une feuille de style XSL (utilisant un script Javascript).

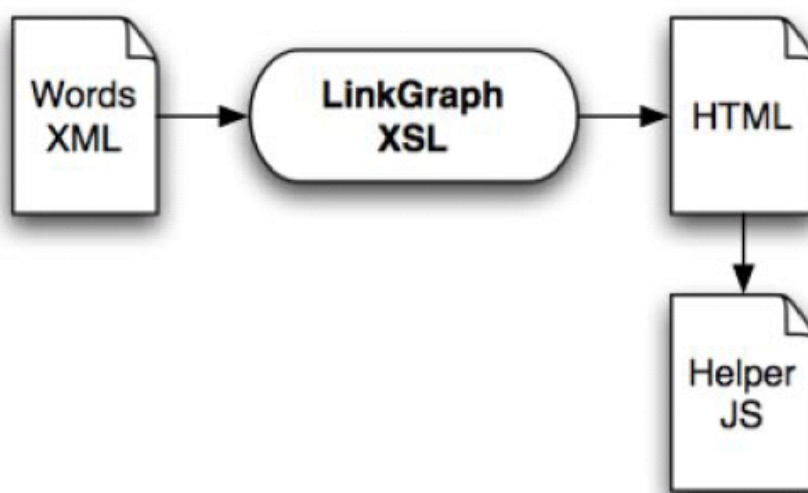


Figure 9 : Architecture "en aval"

Cette architecture « en aval » maintient intégralement le principe présenté dans [Herrington, 2005]. Plusieurs modifications ont cependant été apportées :

- La feuille de style initiale a d'abord été réécrite pour ne produire qu'un nuage de mot sans lien i.e. sans liens vers les contextes originaux contenant les mots scrutés.
- Elle a aussi été modifiée pour affiner les sorties produites (contextes ou carte des sections)

3.6 Fils de Presse

3.6.1 Le Monde

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

Liste des fils RSS du Monde.fr

Pour récupérer facilement les adresses des flux rss, vous pouvez les recopier depuis les boites texte ci-dessous.

XML A la Une	<code>http://www.lemonde.fr/rss/sequence/0,2-3208,1-0,0.xml</code>
XML International	<code>http://www.lemonde.fr/rss/sequence/0,2-3210,1-0,0.xml</code>
XML Europe	<code>http://www.lemonde.fr/rss/sequence/0,2-3214,1-0,0.xml</code>
XML France	<code>http://www.lemonde.fr/rss/sequence/0,2-3224,1-0,0.xml</code>
XML Société	<code>http://www.lemonde.fr/rss/sequence/0,2-3226,1-0,0.xml</code>
XML Régions	<code>http://www.lemonde.fr/rss/sequence/0,2-3228,1-0,0.xml</code>
XML Entreprise	<code>http://www.lemonde.fr/rss/sequence/0,2-3234,1-0,0.xml</code>
XML Médias	<code>http://www.lemonde.fr/rss/sequence/0,2-3236,1-0,0.xml</code>
XML Aujourd'hui	<code>http://www.lemonde.fr/rss/sequence/0,2-3238,1-0,0.xml</code>
XML Sports	<code>http://www.lemonde.fr/rss/sequence/0,2-3242,1-0,0.xml</code>
XML Sciences	<code>http://www.lemonde.fr/rss/sequence/0,2-3244,1-0,0.xml</code>
XML Culture	<code>http://www.lemonde.fr/rss/sequence/0,2-3246,1-0,0.xml</code>

Figure 10 : Les fils du Monde

3.6.2 Libération

Présentation : ¹ <http://www.liberation.fr/page.php?Article=149907>

Que vous soyez webmestre ou utilisateur d'un logiciel agrégateur d'information, Libération.fr met à votre disposition une adresse web pour avoir accès aux dernières nouvelles au format RSS. Certains logiciels permettent en effet de récupérer en temps réel les titres des dernières infos sur une sélection de sites web (par exemple **RssReader** pour Windows ou **NetNewsWire** pour Mac OS X). Vous pouvez dès maintenant ajouter Libération à vos canaux d'information en renseignant l'adresse suivante :

<http://www.liberation.fr/rss.php>

SUR LE MÊME SUJET

- Vous cherchez? Et bien trouvez à présent!

Certains webmestres expérimentés ne manqueront pas d'utiliser cette même adresse pour récupérer les derniers titres de libération sur leur propre site. Les webmestres plus débutants, où ne disposant pas des technologies de syndication RSS sur leur site pourront quand même faire apparaître les derniers titres de Libération.fr sur leurs pages personnelles sous cette forme :



Figure 11 : Le Fil de Libé

3.6.3 Le Figaro

Présentation : <http://www.lefigaro.fr/xml/>

Les différents fils thématiques du Figaro

UNE

FIGARO	XML
FIGARO économie	XML
FIGARO Magazine	XML
Madame FIGARO	XML
FIGARO Entreprises	XML
FIGARO Scope	XML
FIGARO Etudiant	XML

ACTUALITE

International	XML
Europe	XML
Politique	XML
Société	XML
Sciences & Santé	XML
Débats & Opinions	XML
Éducation	XML
Sports	XML

ECONOMIE

Monde - France	XML
Entreprises	XML
High-tech	XML
Médias & Publicité	XML
Votre argent	XML
Décideurs	XML
Finances	XML

ET VOUS.

Culture	XML
Télévision	XML
Auto & Moto	XML
Au masculin	XML
Multimedia	XML

MADAME FIGARO

Bien-être	XML
Beauté	XML
Mode	XML
Cuisine	XML
Déco	XML
Enfants	XML
Psycho	XML
People	XML
Voyages	XML
Luxe	XML
Mariage	XML

FIGARO SCOPE

Restaurants	XML
Cinéma	XML
Arts	XML
Opéra & Danse	XML
Théâtre	XML
Musiques	XML
Style & ville	XML
Enfants	XML
Week-end	XML

FIGARO Etudiant

Dossiers orientation/métier	XML
-----------------------------	-----

Figure 12 : Les Fils du Figaro

4 Liens/projets

4.1 Les nuages de Tags chez Technorati⁹

Principe :

1. *Arranging the words and terms in one paragraph, and*
2. *Varying the font-sizes to represent the popularity of a keyword/ tag.*

Currently tracking 21.7 million sites and 1.7 billion links. [Member Sign In](#) [Sign Up](#) [Help](#)

Technorati™ Search Tags **BETA** Blog Finder Popular About

Search Search Options

Sponsored Links

Nobody reading your blog?
Explode your Blog Traffic. 100% free blog traffic generator.
www.blogexplosion.com

Blogging Evolved
Elegant. Powerful. Professional. The better way to put a blog online
www.squarespace.com/

Car Crazy Community
Upload your car videos and pics join groups, blogs, special events
www.CarCrazyCentral.com

Start your blog now
Publish, be read, and get paid. Start writing instantly!
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 ...

Figure 13 : Nuages de 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.

Plus d'infos : <http://www.technorati.com/help/tags.html>

⁹ <http://www.technorati.com/tags/>

4.2 Annuaire de Fils

LaMooche.fr¹⁰ est un système d'information en perpétuelle évolution qui récupère périodiquement les actualités issues de plus de **1 000 diffuseurs de contenu** (LeMonde, Libération, Le Nouvel Observateur, Clubic, Jeux Video.com, ...). Ce procédé s'appelle l'agrégation de contenu¹¹ (procédé de lecture et de stockage d'articles issus de plusieurs fils d'information).

Annuaire Actualités :

<http://www.lamooche.com/2,1,annuaire-rss-actualite.html>

4.3 AlertInfo, un agrégateur RSS de la presse française

Source : <http://www.geste.fr/alertinfo/home.html>

Communiqué de presse

26 mai 2005

France Télévisions Interactive, Le Monde Interactif, RTL Net, Libération.fr, Les Echos.fr, Le NouvelObs.com, L'Equipe.fr, 01Net, ZDNet, La Tribune.fr, Le Figaro.fr, L'Express.fr, L'Expansion.com, L'Entreprise.com, BusinessMobiles.fr, tous membres du GESTE, lancent, le 26 mai prochain, AlertInfo, lecteur RSS légal et gratuit des médias d'information français, proposant, dès son installation, près de 274 fils d'informations ciblées.

Cette initiative, première mondiale, vise à apporter aux internautes francophones un outil d'information exceptionnel, mis à jour en permanence, promu et réalisé directement par les éditeurs et les rédactions des grands médias électroniques français.

En un seul clic, l'utilisateur aura accès à ses infos préférées : France, International, Business/Entreprises, Communication/Médias/HighTech, Emploi/RH/Métiers, Etudiants/Formations, Solutions IT/Informatique/Matériels, Les Marchés/Investisseurs, Loisirs/Week-end/Culture, Musique, Patrimoine, Sports, Régions, Sciences, Insolite/People, Féminin.

Téléchargeable gratuitement sur le site du GESTE www.geste.fr et sur l'ensemble des sites des éditeurs présents dans le lecteur, AlertInfo permet à chacun de sélectionner ses thématiques et ses sources préférées. Les dernières infos sont présentées par titre et par ordre de mise à jour. Lorsque l'internaute clique sur un titre, le « chapô » se développe dans une partie du lecteur et propose un lien vers le site de l'éditeur pour lire l'article. Pour les articles payants, si l'internaute est abonné, il accède directement à l'article, sans avoir à s'identifier.

Pour être présent dans le lecteur RSS du GESTE, l'éditeur doit être membre du groupement (radio, presse, pure player ou télévision, hors agences de presse), et détenteur des droits des

¹⁰ <http://www.lamooche.fr/>

¹¹ <http://www.lamooche.com/definition/agregation.php>

articles qu'il propose. Les textes proposés doivent être des textes d'information et non des textes de promotion.

Avec AlertInfo, les éditeurs ont souhaité apporter une réponse légale aux attentes des internautes en matière d'information. La réutilisation des fils d'information ainsi mise à la disposition des internautes est soumise aux conditions de chacun des éditeurs.

Le lecteur AlertInfo est le fruit d'un travail d'équipe. Les équipes de développement des echos.fr ont réalisé une version française de la technologie FeedReader (application issue du monde du logiciel libre), en collaboration avec son initiateur, Toomas Toots, à partir de laquelle chacun a apporté sa pierre pour consolider l'édifice (rédaction de l'aide, création d'une charte et de conditions d'utilisation, définition des catégories, etc.).

AlertInfo propose de multiples fonctionnalités, notamment :

- La possibilité d'ajout ou de retrait de fils d'information ;
- La visualisation des « chapôts » (si disponibles) dans une partie de l'écran ;
- la possibilité de trier les fils d'information par catégorie ou par éditeur
- La possibilité d'envoyer l'url d'un article à un ami ;
- et l'envoi d'alertes à chaque nouvelle mise à jour, etc.

Éditeurs présents au lancement d'AlertInfo : lesechos.fr, latribune.fr, lemonde.fr, lefigaro.fr, lentreprise.com, lexpress.fr, lexpansion.com, france2.fr, liberation.fr, rtl.fr, nouvelobs.com, france3.fr, lequipe.fr, 01net.fr, zdnet.fr, businessmobile.fr et, très prochainement, france5.fr et m6.fr

contact

Laure de Lataillade : contact@alertinfo.fr

Astrid Flesch : a.flesch@geste.fr

Tél. 01 55 62 00 70

A propos du Geste :

Le GESTE, qui regroupe les principaux éditeurs de contenus sur internet (presse, radios, télévisions, éditeurs indépendants), a pour objet de créer les conditions économiques, législatives et concurrentielles indispensables au développement des services et éditions électroniques. Avec plus d'une centaine de sociétés membres, le GESTE poursuit sa constante progression et s'est imposé comme l'interlocuteur privilégié et incontournable en matière de contenus en ligne.

Pour télécharger l'outil : <http://www.geste.fr/alertinfo/telecharger.html>

Mode d'emploi : <http://www.geste.fr/alertinfo/modedemploi.html>

4.4 Amazon concordance

Concordance

Concordance is an alphabetized list of the most frequently occurring words in a book, excluding common words such as "of" and "it." The font size of a word is proportional to the number of times it occurs in the book. Hover your mouse over a word to see how many times it occurs, or click on a word to see a list of book excerpts containing that word.

Please send your feedback on this feature to sitb-feedback@amazon.com

Sur le site Amazon.com, présentation du livre : *"In the Beginning...was the Command Line"*, par Neal Stephenson, 1999 :



The screenshot shows the Amazon.com product page for the book "In the Beginning...was the Command Line" by Neal Stephenson. The page features a search bar at the top, navigation links, and a detailed product description. The book cover is prominently displayed on the left, with a "SEARCH INSIDE!" button overlaid. The right side of the page provides the book's title, author, price, and shipping information.

SEARCH INSIDE!™

Neal Stephenson

In the Beginning...was the Command Line (Paperback)

by [Neal Stephenson](#) "Around the time that Jobs, Wozniak, Gates, and Allen were dreaming up these unlikely schemes, I was a teen living in Ames, Iowa..." [\(more\)](#)

SIPs: [bug database](#), [command line interface](#), [window manager](#), [bug report](#)

CAPs: [Was the Command Line](#), [Disney World](#), [Bill Gates](#), [Microsoft Word](#), [Hole Hawk](#) [\(more\)](#)

★★★★☆ [\(85 customer reviews\)](#)

List Price: \$10.00

Price: \$8.00 and eligible for **FREE Super Saver Shipping** on orders over \$25. [See details](#)

You Save: \$2.00 (20%)

Availability: Usually ships within 24 hours. Ships from and sold by Amazon.com.

Amazon Visa® Reward Points: 24
Points are calculated based on the final amount charged.

Want it delivered Tuesday, November 8? Order it in the next 31 hours and 46 minutes, and choose **One-Day Shipping** at checkout. [See details](#)

119 used & new available from \$1.95

Other Editions	List Price	Price	Other Offers
Hardcover	\$18.80		Order it used!

Figure 14 : Amazon "In the Beginning...was the Command Line"¹²

En passant la souris sur l'image de la couverture du livre, on accède au menu suivant :

¹² <http://www.amazon.com/exec/obidos/tg/detail/-/0380815931/002-1510917-1432801?v=glance>

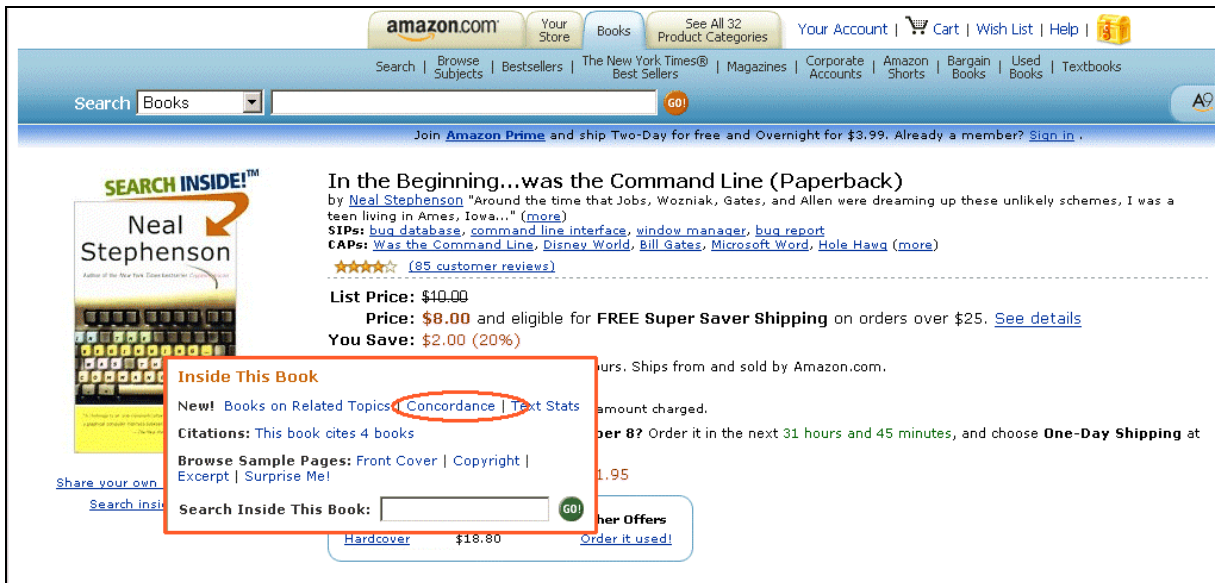


Figure 15 : Menu "concordance" sur Amazon

Ce menu donne accès à un programme « Concordance » qui construit dans un premier temps un nuage de mots (les 100 mots les plus fréquents du livre) :

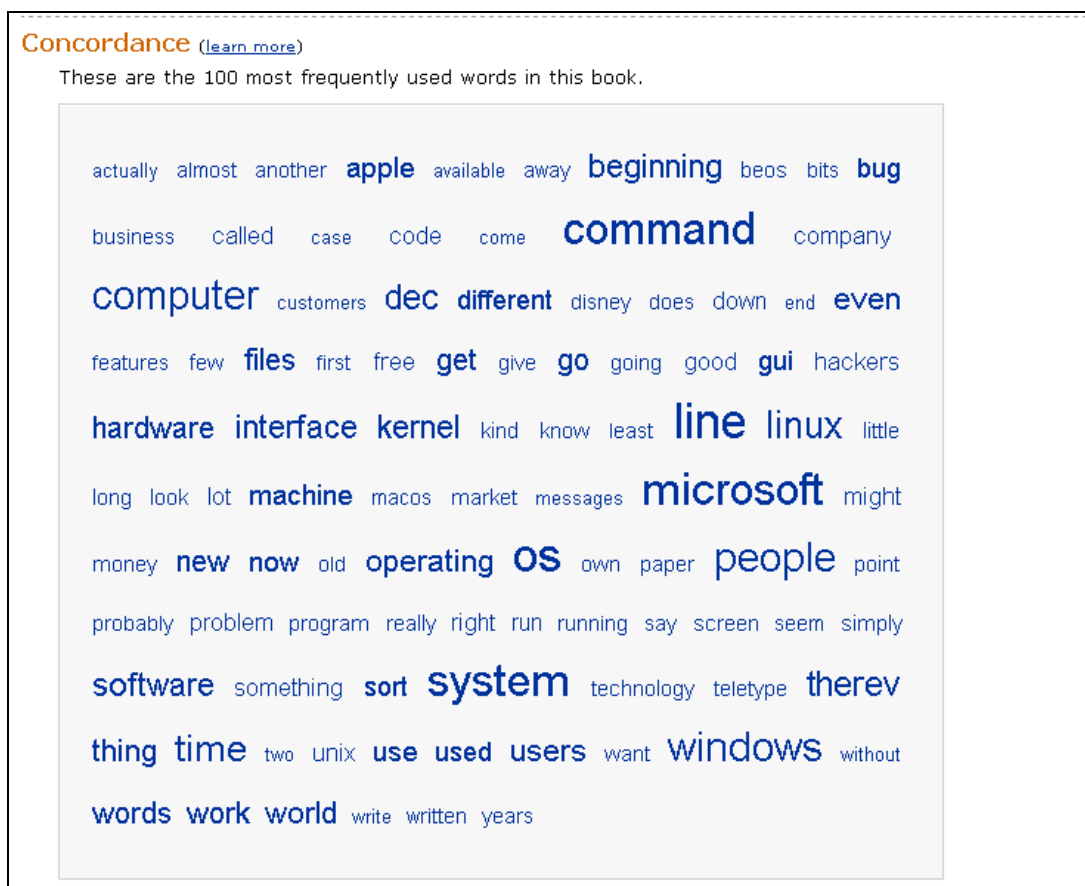


Figure 16 : Nuage de mots "concordance"¹³

¹³ http://www.amazon.com/gp/product/sitb-next/0380815931/ref=sbx_con/002-1510917-1432801?%5Fencoding=UTF8#concordance

chaque mot est ensuite cliquable et donne ainsi accès aux contextes du mot visé :

In the Beginning...was the Command Line
by Neal Stephenson

Price: **\$8.00** [Add to Cart](#)
119 used & new from \$1.95

[View: Front Cover](#) | [Copyright](#) | [Excerpt](#) | [Surprise Me!](#)

81 pages with references to **command in this book:**

- [on Page 3:](#)
"... . . . Was the **Command** Line invented to restrain the power of nineteenth-century robber barons. Item: a woman friend of mine recently told me that ..."
- [on Page 5:](#)
"... . . . Was the **Command** Line The other, somewhat subtler point, was that interface is very important. Sure, the MGB was a lousy car in ..."
- [on Page 7:](#)
"... . . . Was the **Command** Line even than the Euro-sedans, better designed, more technologically advanced, and at least as reliable as anything else on the ..."
- [on Page 11:](#)
"... . . . Was the **Command** Line image of this man sitting there, gripped in the opening stages of an atavistic fight-or-flight reaction, with millions of ..."
- [on Page 13:](#)
"... . . . Was the **Command** Line logics for translating letters into bits and vice versa: teletypes and punch card machines. These embodied two fundamentally different ..."
- [on Page 14:](#)
"... eldritch flavor among those of us who even knew it existed. We were all off the batch, and on the **command** line, interface now-my very first shift in operating system paradigms, if only I'd known it. A huge stack of accordion-fold ..."
- [on Page 17:](#)
"... . . . Was the **Command** Line

	<TR>	
	<TD VALIGN=TOP	
	ROWSPAN="5">	
	</TD>	
	<TD VALIGN=TOP COLSPAN="2"> ..."	
- [on Page 18:](#)
"... how to work with. When we used actual telegraph equipment (teletypes) or their higher-tech substitutes ("glass teletypes," or the MS-DOS **command** line) to work with our computers, we were very close to the bottom of that stack. When we use most ..."
- [on Page 19:](#)
"... . . . Was the **Command** Line systems, though, our interaction with the machine is heavily mediated. Everything we do is interpreted and translated time and ..."
- [on Page 20:](#)
"... it when Microsoft endorsed the idea of GUIs by coming out with the first Windows system . At this point, **command**-line partisans were relegated to the status of silly old grouches, and a new conflict was touched off: between users of ..."

Figure 17 : Contextes "Concordance"

4.5 TagClouds (« Nuage de mots »)

URL du projet : <http://tagcloud.com>

Welcome to TagCloud.com

What is TagCloud?

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.

TagCloud lets you create and manage clouds with content you are interested in, and lets you publish them on your own website.

Sound Interesting?

Lots of other people think so too. The [technology](#) behind TagCloud.com was created just for fun by [IonZoft](#) developer John Herren, and word quickly spread through the blogosphere. After numerous requests for his source code, we decided to produce this service based on John's [original idea](#).

[Sign up](#) for our service absolutely free, or just [learn more](#) about what TagCloud does. Maybe you're interested in [IonZoft](#), the company behind the scenes. If you can't find what you're looking for, please [contact us](#).

What does a TagCloud look like?

It's a list of keywords taken from the news feeds you specify. Larger fonts indicate a higher prevalence for an individual keyword. Using Cascading Style Sheets, you can customize almost every aspect of your TagClouds to make it match your website. Of course, we provide a nice default set of styles out of the box.



amd apparently apple asks autopia big brother blog bloggers browser business model case cell phones chips chris kohler cisco collaborate computer computers cope dalton david discovery donations drm earth engineers exploit fuel google google maps help hybrid intel ipod joanna glasner kohler law enforcement live mail media mice microsoft miles mirrormask missouri mobile mobile phone mozilla firefox open source operating system oregon org partnership peer to peer performance privacy advocates proprietary robotic running scientists search service slashdot socket space spyware state stem cell stem cells story tiny traffic university wifi wired magazine

Figure 18 : projet TagCloud

Application : «TagCloud Le Monde »

Cloud "Le_Monde"

My Clouds View Feeds Edit Stop Word List Import OPML

View Cloud...

Fill RSS



accusations aeqis affaire amiante annonce aot aprs argent arme arrt avant bilan bord breton bush ces cat champion chef chine chinois commerce constitution corruption crise cyclone dbat dcs de france de lutte dominique edf elle emploi encore etats unis europe europen euros exposition france gaza gouvernement grande bretagne grippe aviaire http huit ils irak iran italie japon jean jeu katrina la france la nouvelle la police lance lancer le groupe lemonde loi londres marins matre mis mission mobile new york onu pakistan par paris pdg premier ministre premiere prs publicis relance renault salaris ses sncm soldats street tait tous les travail trs tte ump union victoire villepin wall wall street washington yasukuni york

You can view your public cloud page at http://www.tagcloud.com/cloud/html/Le_Monde/default/50

Figure 19 : tagcloud sur Fils du Monde

Les fils utilisés pour construire ce nuage sont paramétrables via l'onglet Feeds visible dans la figure précédente :

Cloud "Le_Monde"

My Clouds View Feeds Edit Stop Word List Import OPML

Feeds...

Here's where all the fun happens. Enter the URL of the RSS feed you want associated with this cloud, and we will automatically update your cloud with the important keywords.

You can also [specify an OPML File](#) to quickly load multiple feeds.

We update feeds several times a day to make sure your cloud has the most relevant, trendy, and up-to-date information.

We've determined that the best clouds use feeds that have something in common, so you might get strange results if you add feeds about drag racing to a cloud about gardening. That being said, it's your cloud, so do whatever you want!

Feed URL:

Feed Description	
Le Monde.fr : A la Une XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : Aujourd'hui XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : Culture XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : Entreprises XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : France XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : International XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : M?dias XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : R?gions XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : Sciences XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : Soci?t? XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete
Le Monde.fr : Sports XML http://www.lemonde.fr Toute l'actualité? au moment de la connexion	Delete

Figure 20 : Paramétrage du tagcloud LeMonde

4.6 Le filtre Google

<http://www.marumushi.com/apps/newsmap/newsmap.cfm>

Newsmap is an application that visually reflects the constantly changing landscape of the [Google News](#) news aggregator. A treemap visualization algorithm helps display the enormous amount of information gathered by the aggregator. Treemaps are traditionally space-constrained visualizations of information. Newsmap's objective takes that goal a step further and provides a tool to divide information into quickly recognizable bands which, when presented together, reveal underlying patterns in news reporting across cultures and within news segments in constant change around the globe

Newsmap does not pretend to replace the googlenews aggregator. It's objective is to simply demonstrate visually the relationships between data and the unseen patterns in news media. It is not thought to display an unbiased view of the news, on the contrary it is thought to ironically accentuate the bias of it.



Figure 21 : Projet NewsMap

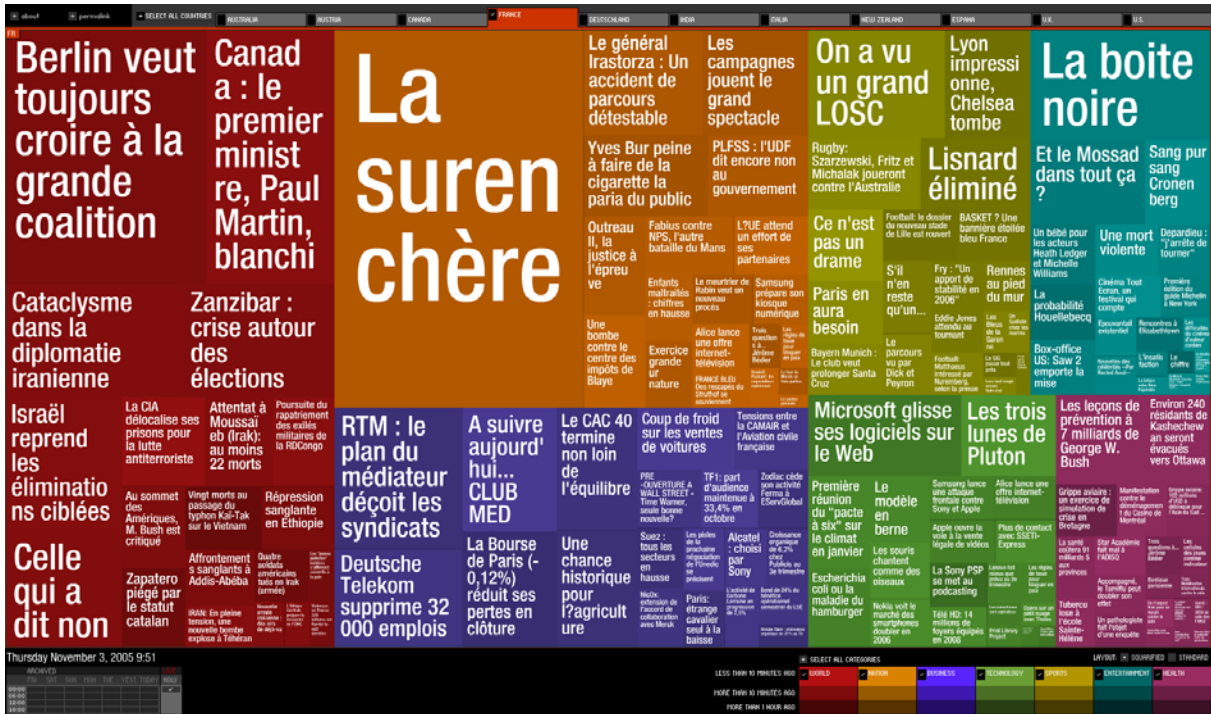


Figure 22 : Projet NewsMap (france)

4.7 10x10 : images du monde

Cet outil d'exploration interactive passe au crible les fils RSS de Reuters, de la BBC et du NewYorker pour créer une mosaïque de photos d'actualités.

Lien : <http://tenbyten.org/index.html>



10 x 10

Process.

Every hour, 10x10 scans the RSS feeds of several leading international news sources, and performs an elaborate process of weighted linguistic analysis on the text contained in their top news stories. After this process, conclusions are automatically drawn about the hour's most important words. The top 100 words are chosen, along with 100 corresponding images, culled from the source news stories. At the end of each day, month, and year, 10x10 looks back through its archives to conclude the top 100 words for the given time period. In this way, a constantly evolving record of our world is formed, based on prominent world events, without any human input.

Sources.

Currently, 10x10 gathers its data from the following news sources:

- ◆ Reuters World News
- ◆ BBC World Edition
- ◆ New York Times International News

10x10

10x10™ ('ten by ten') is an interactive exploration of the words and pictures that define the time.

The result is an often moving, sometimes shocking, occasionally frivolous, but always fitting snapshot of our world. Every hour, 10x10 collects the 100 words and pictures that matter most on a global scale, and presents them as a single image, taken to encapsulate that moment in time. Over the course of days, months, and years, 10x10 leaves a trail of these hourly statements which, stitched together side by side, form a continuous patchwork tapestry of human life.

10x10 is ever-changing, ever-growing, quietly observing the ways in which we live. It records our wars and crises, our triumphs and tragedies, our mistakes and milestones. When we make history, or at least the headlines, 10x10 takes note and remembers.

Each hour is presented as a picture postcard window, composed of 100 different frames, each of which holds the image of a single moment in time. Clicking on a single frame allows us to peer a bit deeper into the story that lies behind the image. In this way, we can dart in and out of the news, understanding both the individual stories and the ways in which they relate to each other.

10x10 runs with no human intervention, autonomously observing what a handful of leading international news sources are saying and showing. 10x10 makes no comment on news media bias, or lack thereof. It has no politics, nor any secret agenda; it simply shows what it finds.

With no human editors and no regulation, 10x10 is open and free, raw and fresh, and consequently a unique way of following world events. In 10x10, we respond instinctively to patterns in the grid, visual indicators of relevance. When we see a frequently repeated image, we know it's important. When we see a picture of a movie star next to a picture of dead bodies, we understand the extremes that exist in our world. Scanning a grid of pictures can be more intuitive than reading headlines, for it lets the news come to life, and everything feels a bit less distant, a bit closer to heart, and maybe, if we're lucky, gives us pause to think. If you'd like to learn more about 10x10, you can read [how it works](#).

Credits.



FABRICA

10x10 was designed and developed by Jonathan Harris of Number27, in conjunction with the FABRICA communication research center in Italy.

Special thanks to: [Andy Cameron](#), [Joel Gethin Lewis](#), [Francesca Granato](#), [David Towey](#)

Figure 23 : Projet 10x10

4.8 Projet « Post Remix¹⁴ » (*Washington Post*¹⁵)

Présentation de ce projet à partir d'un billet publié le 27 novembre 2005 sur le weblog « La feuille¹⁶ » :

Anticiper les usages des lecteurs

<http://lafeuille.blogspot.com/2005/11/anticiper-les-usages-des-lecteurs.htm>

Récemment, en France, le débat vieux médias/nouveaux médias a été relancé, à l'occasion de l'annonce du basculement de Libération vers un modèle de publication "bi-médias", après la refonte récente des maquettes de ses deux grands concurrents nationaux : Le Monde et Le Figaro, explicitement pensées comme repositionnées par rapport à Internet. Une autre manière de présenter l'information, une autre information, une plus grande rapidité, un autre ton, une plus grande interactivité ; c'est toujours sous cet aspect que les médias classiques semblent devoir présenter leur stratégie de communication sur Internet. Il me semble pourtant qu'ils manquent l'essentiel, en considérant toujours leurs lecteurs comme...des lecteurs justement, sans jamais se demander ce qu'ils vont bien pouvoir faire des informations auxquelles ils ont accès.

Donner à d'autres la possibilité de faire quelque chose des informations que l'on publie, les autoriser et leur permettre techniquement de construire des services à partir d'un flux d'informations et de remixer ce flux pour offrir des contenus recomposés, c'est ce que vient de faire le Washington Post en ouvrant son service "Post remix¹⁷". Il s'agit ni plus ni moins de permettre à quiconque de programmer des mashups du Washington Post sur la base d'API fournies par le journal. Un premier mashup permet de créer un flux RSS sur les résultats de recherche¹⁸ par mots-clés sur les résultats du Post, et un autre, très intéressant, représente ces mots-clés en nuage de tags¹⁹, à la Del.icio.us.

C'est assez futé tout de même, car le journal externalise ainsi à bon compte sur sa communauté de lecteurs tous les services qu'il aurait pu développer lui-même. Par ailleurs, la licence d'utilisation²⁰ vaut le coup d'être lue :

Your efforts must be for personal, and not for commercial, use. You may not sell applications that use or incorporate washingtonpost.com content.

You recognize that Washingtonpost.Newsweek Interactive retains all intellectual property rights in all washingtonpost.com content and you that acquire no such rights by participating in Post Remix.

Washingtonpost.com may incorporate your ideas into future projects it develops.

¹⁴ http://blogs.washingtonpost.com/post_remix/

¹⁵ <http://www.washingtonpost.com/?nav=globaltop>

¹⁶ <http://lafeuille.blogspot.com/>

¹⁷ cf supra

¹⁸ <http://socialistsushi.com/wp/>

¹⁹ <http://www.revsys.com/newscloud/>

²⁰ http://blogs.washingtonpost.com/post_remix/2005/11/terms_of_use.html

Appréciations en particulier la dernière clause, assez savoureuse dans le style faut-pas-se-gêner... C'est d'ailleurs tout le problème des remix²¹, qui reposent le plus souvent sur des bases de coopération pas claires et totalement déséquilibrées.

Au delà d'un rapport de force qui devra nécessairement s'équilibrer, l'exemple est quand même inéressant : voilà un grand journal qui commence à repenser sa position dans la chaîne de circulation de l'information et considère davantage ses lecteurs comme des partenaires. A méditer.

L'application « nuage de tags » présentée dans ce billet (**NewsCloud**) donne à voir un processus similaire à celui mis en œuvre dans le projet présenté dans ce document. La page d'accueil du site Newscloud est présentée ci-dessous :

The screenshot shows the NewsCloud interface for the keyword 'bush'. On the left, there is a list of article titles and snippets, including 'Transcript of President Bush's Press Conference', 'Bush's Tortured Logic', 'Cheney's Challenge', 'The Trust is Gone', and 'White House Gambles That Boosting Kilgore Will Pay Off for Bush'. On the right, there is a 'ZOOM' button and a tag cloud containing various terms like 'iraq', 'white house', 'supreme court', 'washington', 'bush', 'united states', 'chicago', 'fbi', 'parents', 'fitzgerald', 'baltimore', 'south', 'fairfax', 'republicans', 'family', 'cheney', 'america', 'help', 'johnson', 'republican', 'wilson', 'price', 'dia', 'indictment', 'china', 'york', 'money', 'ipod', 'media', 'lead', 'local', 'kilgore', 'university', 'george', 'games', 'space', 'flu', 'japan', 'job', 'ball', 'maryland', 'white house', 'search', 'afghanistan', 'congress', 'hurricane', 'katrina', 'oil', 'federal', 'miers', 'new orleans', 'troops', 'ravens', 'united', 'smith', 'american', 'google', 'virginia', 'war', 'prince', 'microsoft', 'senate', 'eagles', 'texas', 'montgomery', 'mail', 'coach', 'williams', 'fire', 'white', 'miller', 'air', 'state', 'pandemic', 'case', 'street', 'president', 'bush', 'hurricane', 'http', 'touchdown', 'united', 'states', 'supreme', 'court', 'iran', 'al qaeda', 'orleans', 'israel', 'washington', 'nomination', 'bush', 'center', 'union', 'new york', 'story', 'wizards', 'game', 'computer', 'democrats', 'victory', 'community', 'libby', 'apple', 'service', 'death', 'syna'.

Figure 24 : Projet NewsCloud (Washington Post)

²¹ <http://www.readwriteweb.com/archives/002829.php>

About NewsCloud

[NewsCloud](#) is an application that takes all of the RSS feeds from the [Washington Post](#) website and builds a blog like [tag cloud](#) from the keywords. Each story's full text is pulled from the website and indexed by keywords thses keywords. There are typically around 11,000 news stories and 60,000 keywords being indexed at any given time.

How to use NewsCloud

When you first go to [NewsCloud](#) you are seeing the outer most zoom of the cloud. The outer most level is where the most popular keywords are. The farther you zoom into the cloud the frequency of the keywords is reduced. You can zoom by clicking the arrow to the right of the big **ZOOM** in the upper right hand corner. Zooming to find less frequent keywords can reveal some interesting topics just below the surface.

At each zoom level, including the outer most, you will see keywords that are in **red**. This is the most frequent keyword at the zoom level you are currently on. As you zoom the stories on the left change. These stories are the ones that contain the keyword in **red**.

To view the articles associated with any keyword on the page simply click on the keyword and the articles will be shown on the left.

Technology Used

[NewsCloud](#) was written by [Frank Wiles](#) as an experiment. It uses a slightly non-standard LAMP like environment. Typical LAMP application use Linux, Apache, MySQL, and a "P" lanugage such as Perl, PHP, or Python. The following technologies are used in NewsCloud:

- [Linux](#)
- [Apache](#)
- [mod_perl](#)
- [PostgreSQL](#)

This slightly different LAMP stack is the preferred development environment of Revolution Systems. Please visit the respective technology pages or [contact](#) us for more information about how your business can benefit from these Open Source technologies.

4.9 1000Tags

- **1000Tags** : [1000tags.com](#) is a project that aims to put to the test in its simplest form the viability of tagging as a way to advertise, by presenting a tag cloud formed by tags added by people who try to promote a particular site or page. This is done by offering a web page where anyone can book a particular tag that will later be displayed in the main tag cloud at the [1000tags.com](#) page, as well as allowing web site owners to add their tag - for free of course - by syndicating a small tag cloud at their pages. <http://1000tags.com/>.

4.10 ZoomTags

- **ZoomTags** : *ZoomTags is a professional implementation of the commercial tagcloud idea introduced by 1000Tags.* <http://www.zoomtags.com/>. (Présentation de ce projet sur le blog "[TechCrunch](#)" dans ce [billet](#)).

5 Lectures

5.1 Conversation : De la représentation visuelle à la complexité documentaire

Source : http://affordance.typepad.com/mon_weblog/2005/11/de_la_representa.html

A moins que ce ne soit l'inverse : De la représentation documentaire à la complexité visuelle. Le site Visual Complexity²² propose, classés par thèmes (biologie, arts, réseaux sociaux, web...), des projets (231 au total) de visualisation de masses complexes d'informations et/ou de documents. Au-delà du côté simplement esthétique de ces représentations de la complexité, au-delà également de l'enjeu technique et (parfois) algorithmique que ces mêmes représentations supposent, elles illustrent parfaitement ce qu'est le principe de "tertiarisation documentaire" : après les documents primaires (ouvrages 'originaux'), après les documents secondaires (ouvrages décrivant le contenu des premiers), voici - au même titre que les cartes heuristiques - les documents tertiaires qui font sens en eux-mêmes (dans la mesure ou ils offrent leurs propres parcours interprétatifs) et/mais n'existent que parce qu'il renvoient (au sens propre et non 'intertextuel') vers d'autres. Ce nouveau genre documentaire m'avait frappé lors de ma découverte des premières cartes du métamoteur Kartoo²³ (on a les illuminations qu'on peut, n'est pas Claudel qui veut ...). Il avait aussi frappé mon collègue Gabriel Gallezot (évoquant un 'Darwinisme documentaire'), à tel point que nous avons commis quelques articles effleurant le sujet. Et le gars Roger, il en pense quoi ?

(Ndt : Roger = Roger Pedauque

cf http://rtp-doc.enssib.fr/rubrique.php3?id_rubrique=13)

5.2 Blog *Technologies du Langage*²⁴ (par Jean Véronis)

Plusieurs billets sont consacrés à une thématique proche.

Web : Surfez sur les nuages

<http://aixtal.blogspot.com/2006/01/web-surfez-sur-les-nuages.html>

Texte : Nuages dynamiques

<http://aixtal.blogspot.com/2005/11/texte-nuages-dynamiques.html>

Texte: Chirac sur un nuage

<http://aixtal.blogspot.com/2005/11/texte-chirac-sur-un-nuage.html>

Blogs: Banlieues dans les nuages

<http://aixtal.blogspot.com/2005/11/blogs-banlieues-dans-les-nuages.html>

²² <http://www.visualcomplexity.com/vc/>

²³ <http://www.kartoo.fr/>

²⁴ <http://aixtal.blogspot.com/>

Blogs: Un nuage sur les banlieues

<http://aixtal.blogspot.com/2005/11/blogs-un-nuage-sur-les-banlieues.html>

Dialogue entre blogues

<http://aixtal.blogspot.com/2005/10/rcr-dialogue-entre-blogues.html>

5.3 Bibliothèque 2.0²⁵

Article paru sur le blog *bibliosession*²⁶ :

On parle souvent de [Web 2.0](#)²⁷, et de toutes les usages sociaux qui l'accompagnent. Fred Cazzava fait d'ailleurs [le point sur ce sujet](#)²⁸ en prenant pas mal d'exemples et surtout en recentrant le débat sur les usages. Ce qu'il y a peut-être de plus important à comprendre peut se résumer dans ce qui était cité par [Hubert Guillaud](#)²⁹ qui citait lui-même [Cyberlibris blog](#)³⁰. En Substance: "On oublie trop souvent l'utilisateur et on ne se préoccupe que du livre, de ses ayants-droit (les maisons d'édition) et de ceux qui aimeraient accéder (par des moyens pas toujours orthodoxes) au copyright des ayants-droit (Google et les autres). Mais, où est donc passé l'utilisateur. Est-il si peu important qu'il n'y a rien à en dire? Je pense qu'il y a là une erreur de perspective fondamentale. La bataille de l'émancipation de la musique et de l'image a été gagnée par les utilisateurs (et les pressions plus ou moins hardies qu'ils ont exercées). Il en va de même du livre. Lorsque l'on interroge les utilisateurs (ce que nous avons fait), que souhaitent-ils vraiment à propos du livre? Trois choses principales:

- **Pertinence:** L'utilisateur veut pouvoir accéder aux livres dont il a besoin. Malheureusement, cette demande est loin d'être satisfaite par les circuits existants. Une librairie, si vaste soit-elle, ne peut stocker tous les livres. Très souvent, elle ne stocke que ce qui se vend. Pour passer des journées entières dans les catalogues d'éditeurs, je suis tout à la fois admiratif de la richesse de l'esprit humain et consterné que si peu en soit visible.

- **Immédiateté:** L'utilisateur a besoin du contenu "maintenant", c'est-à-dire au moment où son besoin d'information s'exprime. Il ne s'agit pas d'avoir une réponse demain. L'utilisateur est prêt à payer cette instantanéité de réponse.

- **Ubiquité:** L'utilisateur souhaite obtenir une réponse à ses besoins d'information où qu'il se trouve. Il est prêt à payer cette ubiquité documentaire.

Si l'on rassemble ces trois exigences à l'instar d'un portrait chinois, on découvre le format approprié à les satisfaire: il s'agit d'une **bibliothèque digitale.**"

A quoi peut donc ressembler cette bibliothèque digitale?

Et bien [cet article](#)³¹ cité par l'excellent blog [Librarian in Black](#)³² explique que suite à [l'Internet Librarian Conférence](#)³³ qui s'est tenue récemment aux Etats-Unis, un groupe d'une centaine de bibliothécaires a souhaité se réunir pour réfléchir à la prise en compte des usages permis par le Web 2.0 dans les bibliothèques. Ce groupe s'est désigné tout logiquement **library 2.0**.

They hope that the Library 2.0 "movement" will break librarians out of brick-and-mortar establishments and get them to interact with patrons through blog comments, IM and Wiki entries.

Alors qu'est-ce que ça donne? Un des exemples est le site de la bibliothèque de la [Thomas Ford Memorial](#)

²⁵ <http://bibliobsession.over-blog.com/article-1246121.html>

²⁶ <http://bibliobsession.over-blog.com/>

²⁷ http://fr.wikipedia.org/wiki/Web_2.0

²⁸ <http://www.fredcavazza.net/index.php?2005/11/20/951-web-20-le-putsch-des-utilisateurs>

²⁹ <http://lafeuille.blogspot.com/2005/11/bibliothèques-numériques-pertinence.html>

³⁰ http://cyberlibris.typepad.com/blog/2005/11/je_viens_de_lir.html

³¹ <http://www.publish.com/article2/0,1895,1881893,00.asp>

³² <http://librarianinblack.typepad.com/librarianinblack/>

³³ <http://www.infotoday.com/il2005/>

[library](#)³⁴ qui se veut "orienté usager" et qui n'a rien de spectaculaire si ce n'est sa grande clarté et l'intégration d'un blog ainsi que la possibilité d'agrandir les caractères. Plus de détails [ici](#)³⁵ et [là](#)³⁶. Il est également intéressant de voir combien les bibliothécaires américains commencent à intégrer les logiciels de [messageries instantanées](#)³⁷ comme moyen de communication avec leur usagers.

D'autres innovations plus spectaculaires permettent d'utiliser des [Tags pour visualiser des collections de bibliothèques](#).

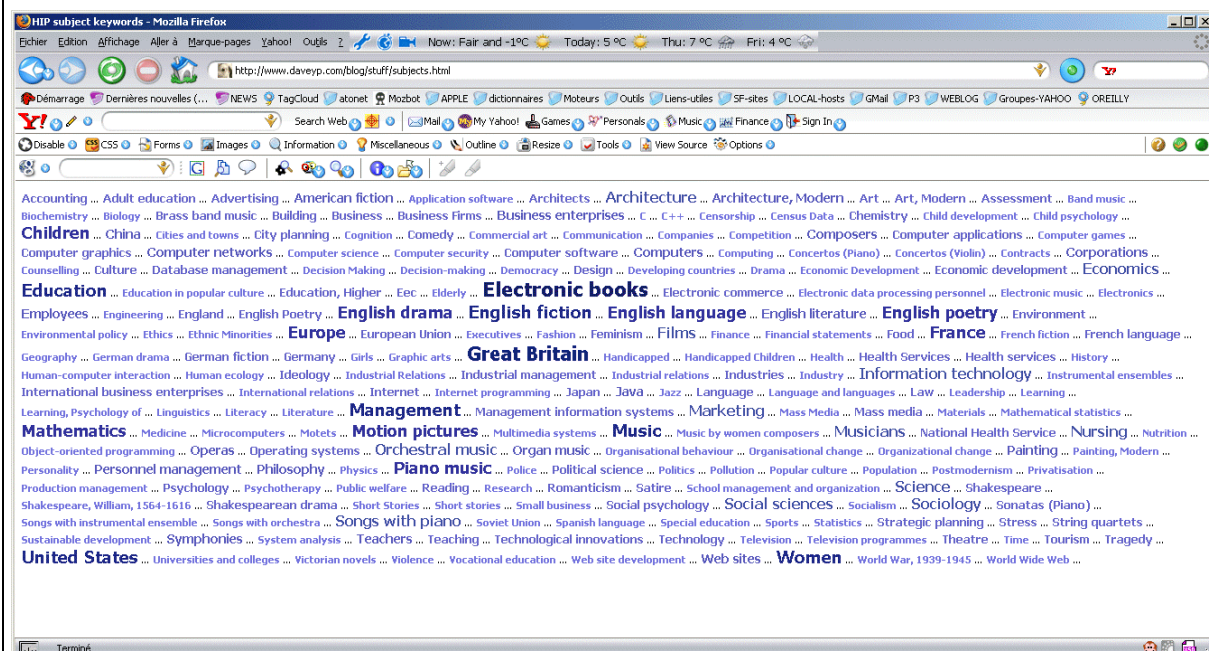


Figure 25 : des Tags pour visualiser des collections de bibliothèques

Ces tags ne relèvent pas du chaos de l'indexation "incontrôlée" (l'absence d'autorités-matières comme sur [del.icio.us](#)³⁸ et bien d'autres). Non ces Tags là sont une manière nouvelle de porter un regard sur la collection et sont reliés à une indexation matière tout ce qu'il y a de plus traditionnel. L'idée a été reprise à partir du travail fait par [Jenny Levine](#)³⁹ et son "prototype" [Mock up](#)⁴⁰. L'auteur de cette expérimentation propose même le [script](#)⁴¹ pour tout ceux qui ont le SIGB Horizon/HIP.

Coté innovations toujours, cet exemple [d'interface d'interrogation](#)⁴² du catalogue réalisé par Casey Bisson de l'université de Plymouth (son blog [ici](#)⁴³) qui fonctionne comme [google suggest](#)⁴⁴. Toutes ces expérimentations préfigurent ce que pourront proposer les futures bibliothèques digitales, à condition que tout cela soit accompagné d'une réflexion sur les pratiques et les usages, avant les prouesses technologiques. Pour finir, ne manquez pas [cet article de Tim O'Reilly](#)⁴⁵ et celui de [Paul Miller](#)⁴⁶ cité sur le blog [Culture et TIC](#)⁴⁷...

³⁴ <http://www.fordlibrary.org/foundation/>

³⁵ <http://www.flickr.com/photos/aaronschmidt/64966024/in/photostream/>

³⁶ <http://www.walkingpaper.org/>

³⁷ <http://walkingpaper.org/181>

³⁸ <http://del.icio.us/>

³⁹

[http://www.theshiftedlibrarian.com/archives/2005/11/06/anybody going to blog these library 20 events.html](http://www.theshiftedlibrarian.com/archives/2005/11/06/anybody_going_to_blog_these_library_20_events.html)

⁴⁰ <http://flickr.com/photos/shifted/60728682/>

⁴¹ <http://www.daveyp.com/blog/index.php/archives/47/>

⁴² <http://www.plymouth.edu/library/bibinfo/suggest.html>

⁴³ <http://www.maisonbisson.com/blog/>

⁴⁴ <http://www.google.com/webhp?complete=1&hl=en>

⁴⁵ <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>

6 Les développements à construire

6.1 Traitements des contenus des fils

Etiquetage, repérage de syntagme, segments répétés etc.

6.2 Archivages des fils de presse (*in-progress*)

Mise en place d'un processus (expérimental) automatique d'archivage des fils RSS du journal Le Monde.

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 (la périodicité reste en fait à déterminer...) (le Monde *EN SURFACE*)

En archivant régulièrement les fils on aura donc à portée de main le *PROFOND* et la *SURFACE*.

Comment ça marche :

- un script aspire les fils en question et crée des pages de navigation...
- le script est activé automatiquement toutes les heures sur la serveur où est installé le script

Les données actuellement sont visibles ici : <http://sfmac.no-ip.com/fils-presse-archivage/index.xml>

(accès restreint pour le moment).

Après validation, ce processus sera mis en place sur le serveur de Paris 3 accompagné des scripts de traitements à mettre en place sur ces fils archivés (les nuages de mots et plus sérieusement des traitements statistiques)

6.3 En Vrac

6.3.1 Commentaires AS

Construire une méthodologie "propre" de traitement de ces données sur des fils ou des mémoires plus anciens, (ou se placer à une date récente avec une mémoire antérieure. Ce sera moins "actuel" mais ça n'est pas trop grave. (Propre= totalement explicite et sans bricolages mal justifiés.)

Quelque pistes :

a) pour juger de l'originalité d'un billet B1, il faut pouvoir :

a1) disposer d'éléments de comparaison

ex : une mémoire globale (un ensemble de fils similaires sur une période / ou les articles du Monde / + etc.)

⁴⁶ <http://www.ariadne.ac.uk/issue45/miller/>

⁴⁷ <http://culturetic.canalblog.com/archives/2005/11/22/1025082.html>

a2) si possible d'éléments pertinents (i.e. en rapport avec B1) donc, extraire de cette "mémoire" un sous ensemble d'unités "probablement en rapport" avec le billet (classification automatique, extraction d'après la présence de certaines unités lexicales, etc.)

a3) appliquer, par exemple, des spécificités chronologiques calculées, bien entendu (pour moi en tous cas) sur les segments et les cooccurrences) pour comparer le billet et la sélection faite à partir de la mémoire

b) Il devrait être utile de s'intéresser à l'histoire et à la localisation "spatiale" de certaines unités, à la fois dans la mémoire totale et dans la mémoire sélectionnée

7 Liens et développements autour de RSS

RSS et la publication simultanée dans Internet

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

par Olivier Charbonneau

Le concept de diffusion simultanée ou syndication n'est pas nouveau. Dès le milieu du 19e siècle, certains grands quotidiens aux États-Unis employaient des mécanismes de diffusion simultanée grâce aux technologies disponibles à l'époque. Entre autres, ils offraient aux quotidiens régionaux des feuilles «pré-imprimés» où figuraient des articles plus pérennes, des annonces ainsi que des illustrations satyriques. C'est ainsi que les bandes dessinées de la grande presse Nord-Américaine pris son essor...

RSS bidirectionnel

http://affordance.typepad.com/mon_weblog/2005/11/rss_bidirection.html

(L'utilisateur peut ne plus "simplement" se contenter de recevoir des informations mais en ajouter et/ou modifier directement le flux. En anglais dans le texte cela s'appelle "bidirectional, asynchronous replication".)

Feed for Thought

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

« *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)