



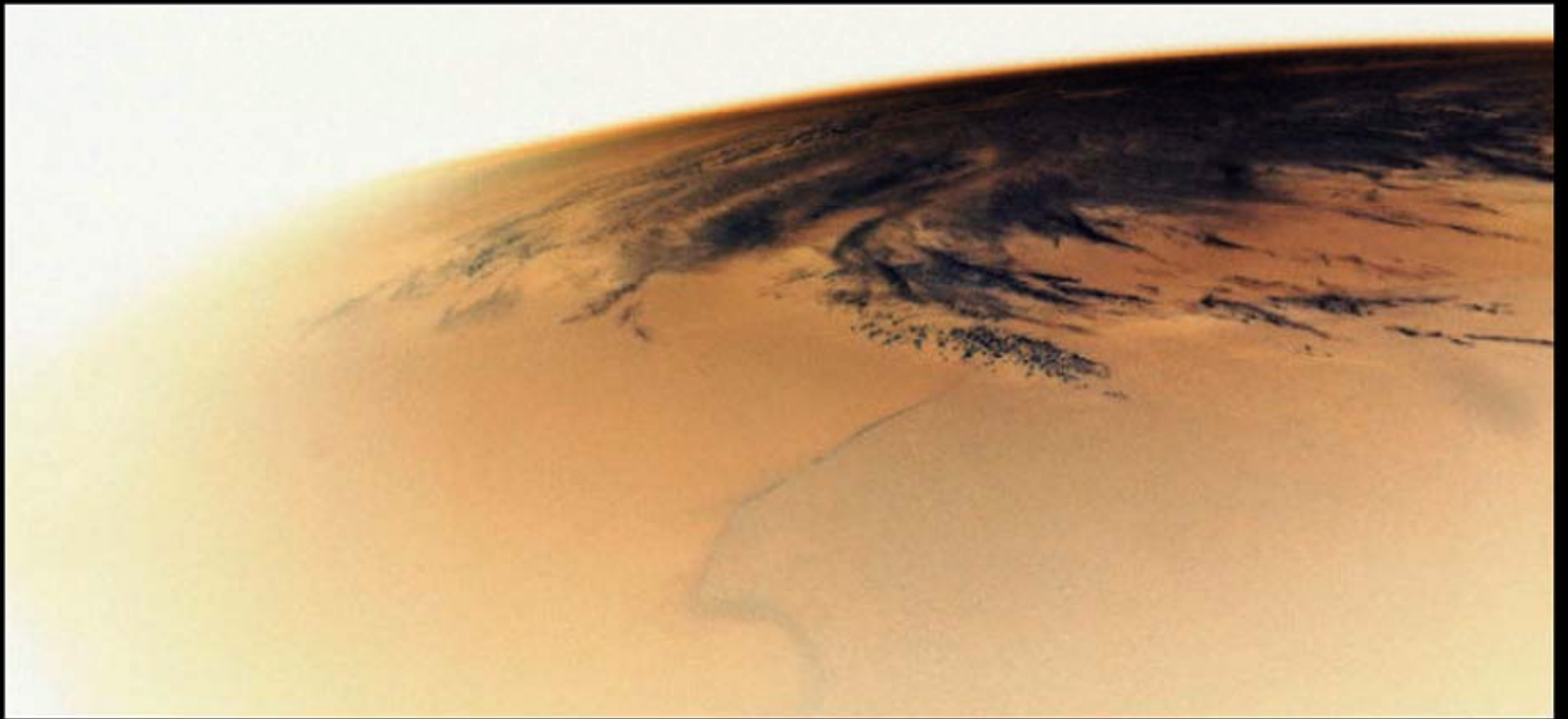
Programmation et projet encadré

J-M. Daube, R. Belmouhoub (CRIM/Inalco), S. Fleury (CLA²T, U. Paris3)



Sommaire

- Préambule
 - Contact et liens
- Le Projet :
 - Présentation générale du projet
 - Les outils, les méthodes, les traces et les rendus
- Projet n°1 **Phase 1**
 - « *échauffement* » « *à la main* » !!!!!
- Projet n°1 **Phase 2**
 - Automatisations (en 3 étapes)
- Les séances « BOITES A OUTILS »
 - *Série 1* : Perl
 - *Série 2* : étiquetage
 - *Série 3* : Extraction terminologique
 - *Série 4* : Des textes aux Graphes (via Pajek)



- Préambule

Site PluriTAL <http://plurital.org/>

PluriTAL

[Accueil](#)
[Actualités](#)
[Administration](#)
[Master PLURITAL](#)
[Journées-Cours](#)
[Liens](#)
[Contact](#)
[Bibliographie](#)
[Lectures](#)
[Groupe PluriTAL](#)
[Images](#)

Filières TAL et ingénierie linguistique de Paris III Sorbonne nouvelle, Paris X Nanterre, INALCO (Institut National des langues et civilisations orientales)

Lieux/Sites PluriTAL

Localisations de Plurital (comment y aller)

Site Web officiel : <http://tal.univ-paris3.fr/plurital/>

Site Web miroir : <http://www.cavi.univ-paris3.fr/ilpga/plurital/>

Rentrée 2005-2006

Journée d'accueil du **MASTER** le 10 octobre 2005, 14h00-16h00, dans les salons de l'INALCO.

PluriTAL - Apports croisés, complémentaires et pluriels pour le TAL

Le domaine du TAL et de l'ingénierie linguistique se caractérise par la multiplicité des dimensions prises en compte (traitement de l'écrit, de l'oral), des niveaux d'analyse impliqués (morphologie, syntaxe, sémantique, pragmatique), des techniques, des langues abordées. C'est un domaine où il est essentiel de conjuguer des apports multiples.

Nous enseignons, effectuons des recherches et participons à des projets dans le domaine du traitement automatique des langues et de l'ingénierie linguistique. Nous travaillons ensemble pour l'enseignement comme pour la recherche.

Nous souhaitons aller plus loin en croisant progressivement nos apports spécifiques pour offrir une formation plus complète. Nous espérons d'ailleurs pouvoir progressivement travailler avec d'autres formateurs du domaine sur la région parisienne.

Nous travaillons depuis plusieurs années universitaires à conjoindre nos formations. Les deux premières années, ce sont des conférences croisées qui ont été mises en place. Nous avons organisé, en 2002-2003, deux matinées de présentation mutuelle des travaux et recherches que nous effectuons, puis en 2003-2004 une **matinée** pour continuer cet échange et préparer les "échanges" de cours et les compléments de formation à intégrer progressivement dans les formations que nous assurons.

De manière plus ambitieuse, nous avons rajouté aux cursus de nos étudiants pour l'**année 2004/2005**, 24 heures de cours assurées (en dehors des services statutaires) à raison de 8h par établissement, sur trois thématiques : la formalisation en syntaxe ; la constitution et l'utilisation de corpus électroniques à des fins lexicographiques ; l'utilisation de ressources électroniques multilingues.

Groupe PluriTAL

<http://tal.univ-paris3.fr/plurital/groupeplurital.html>

[Lectures](#)

- MoDyCo : Laboratoire Modèles, Dynamiques, Corpus. Projets TAL associés : PETAL, METAL

[Groupe PluriTAL](#)

LIMSI

- LIR "*Langues, Information et Représentations*", LIMSI

[Images](#)

INALCO

- INALCO, <http://www.inalco.fr/>.
- Centre de Recherche en Ingénierie Multilingue, <http://www.crim.fr/>.

Groupe Yahoo PluriTAL

Inscription

Inscription ici



S'inscrire à PluriTAL

entrez ici votre adresse 

Adresse du service :fr.groups.yahoo.com

Accès au site du groupe **pluriTAL-yahoo**

<http://fr.groups.yahoo.com/group/plurital/>

Réactions - Commentaires - Dialogues croisés...

[A lire sur cette page](#)

Page web du cours

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

- [Accueil](#)
- [Nouveautés](#)
- [Sommaire](#)
- [Cours en ligne](#)
- [Enseignements](#)
- [Parcours TAL](#)
- [Lectures](#)
- [Liens](#)
- [PluriTAL](#)
- [TALetTOILE](#)
- [Corpus](#)
- [Logiciels](#)
- [Etudiants TAL/ILPGA](#)
- [Travaux Etudiants](#)
- [Mémoires IDL](#)
- [Examen/Partiel](#)

Rechercher sur ce site avec [Antinéa](#) :

Secteur TAL Informatique

ILPGA Université Paris 3

Cours Modules L7T04/L8T07 "Programmation et projet encadré" (Master TAL 1/2) TAL Paris 3 - ILPGA / Paris X / INALCO

Responsables du cours

S. Fleury/B. Habert/J.M Daube. serge.fleury@univ-paris3.fr

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.

Ressources pour le cours

Contrôle des connaissances

Bibliographie

Liens

[Portail TAL - P3](#) : cours, Tds, outils...

[Claroline](#) : ressources complémentaires pour le cours

PLURITAL : Filières TAL et ingénierie linguistique de Paris III Sorbonne nouvelle, Paris X Nanterre, INALCO (Institut National des langues et civilisations orientales)- Apports croisés, complémentaires et pluriels pour le TAL

Contacts et hypertoile

- Courrier électronique
 - serge.fleury@univ-paris3.fr
 - rbelmouhoub@inalco.fr
 - Jmdaube@inalco.fr
- Site *pluriTAL*
 - <http://plurital.org>
- Groupe *plurital-yahoo*
 - <http://fr.groups.yahoo.com/group/plurital/>
- Blog *pluriTAL*
 - <http://tal-p3.wordpress.com>



Le projet

- [Présentation générale](#)

De la théorie...

• Descriptif du cours

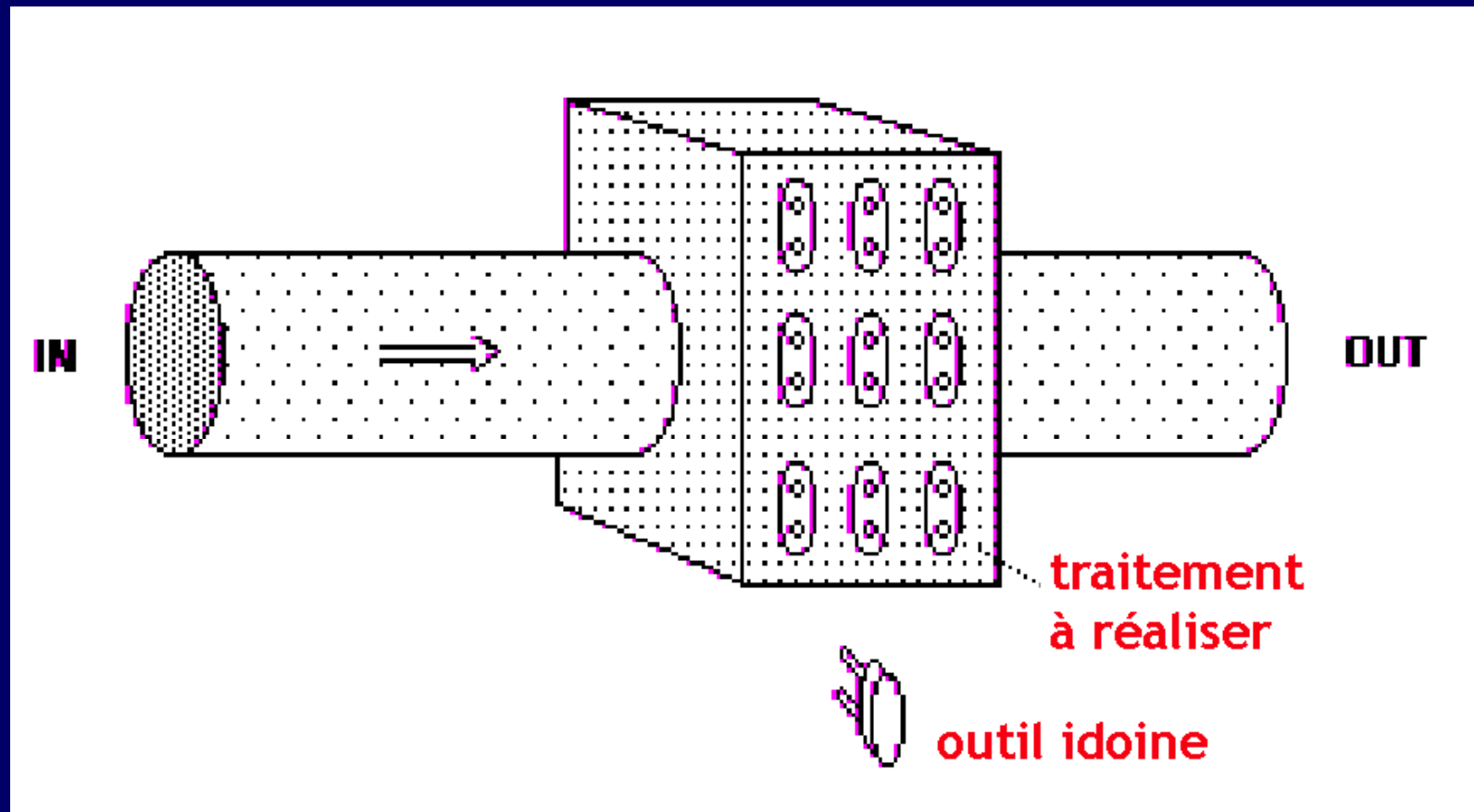
Mise en oeuvre d'une **chaîne de traitement textuel semi-automatique**, depuis la récupération des données jusqu'à leur présentation.

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 dessous des cartes

- *IN* : données fournies + objectif
- Des traitements (*séquencés par nous*) à réaliser (*par vous*) sur les données avec des outils adéquats (*qu'il faudra apprendre à connaître*)
- *OUT* : données à nous remettre
 - (rapport + programme)



Le projet

- Des exemples de projet réalisés les années passées

Le dessous des cartes: Projets 2005-2008

- Navigations dans Les Fils du Monde 07/08
 - <http://tal-p3.wordpress.com/navigation-dans-les-fils-du-monde-0708/>
- Projet *Multilinguisme* 07/08
 - <http://tal-p3.wordpress.com/multilinguisme-0708/>
- Navigations dans Le Monde 06/07
 - <http://tal-p3.wordpress.com/page-navigations-dans-le-monde>
- Projet *Nuages* 05-06
 - <http://tal-p3.wordpress.com/page-nuages/>
- Projet *Multilinguisme* 05-06
 - <http://tal-p3.wordpress.com/page-multilinguisme/>
- Projet *Communauté* 05-06
 - <http://tal-p3.wordpress.com/page-communaute/>
- Voir aussi sur le blog pluriTAL : pages des Projets
 - <http://tal-p3.wordpress.com/>

Le dessous des cartes (EXEMPLE 1 ^{1/4})

- *Cf Cours n°2 PLURITAL 2004/2005*
 - <http://tal.univ-paris3.fr/plurital/cours2-2004.html>
 - **Ajuster corpus et objectifs**
 - Objectif :
 - décrire un mot/une notion **sur** corpus (laïcité) dans Le Monde

Le dessous des cartes (EXEMPLE 1 ^{2/4})

- *IN* : Le journal « Le Monde »
 - Période n°1 : Avril 2003 - Octobre 2004
 - Plusieurs rubriques regardées (environ 5000 articles (pour le CORPUS FRANCE qui regroupe les rubriques "France" et "France-Société")
 - Période n°2 : 1999-2000
 - Ce corpus couvre la période 1991-2000 et ne comporte que les documents contenant au moins un des mots de la famille "laïcité". Il compte 2837 articles, 3 000 000 de mots environ (et pèse 17 Mo environ). Ce corpus est disponible ici en 2 versions : **Texte** et **Lemmatisé**. Pour chacune des ces versions, on dispose des indications suivantes sur chaque article : ANNEE, MOIS, RUBRIQUE et GENRE

Le dessous des cartes (EXEMPLE 1 ^{3/4})

- Traitements des données :

- Période 2003-2004

- Aspiration des pages (HTTRACK)
 - Etiquetage (CORDIAL)
 - Normalisation (SCRIPTS PERL)
 - Concordances (LEXICO)
 - Graphiques de ventilation (LEXICO)

- Période 1999-2000

- *« En marge de projets de recherche, une base documentaire, nommée LM10, a été constituée à partir des archives électroniques de 10 années du journal Le Monde : 1991 à 2000. Cette base totalise environ 200 millions de mots (à titre d'ordre de grandeur, rappelons qu'un roman de taille moyenne comprend de l'ordre de 100 000 mots. LM10 correspond à quelque 2 000 romans)... »*
(suite)

Le dessous des cartes (EXEMPLE 1 ^{4/4})

- *OUT* : données à nous remettre
 - (rapport + programme)
 - http://tal.univ-paris3.fr/plurital/cours2-2004/Ajuster_Corpus_Recherche_Cours_2.html
 - http://tal.univ-paris3.fr/plurital/cours2-2004/Ajuster_Corpus_Recherche_Cours_3.html

Le dessous des cartes (EXEMPLE 2 ^{1/3})

Construction d'un lexique bilingue des médias

IN (français)

<http://www.pch.gc.ca/progs/ac-ca/progs/ri-bpi/pubs/juneau/francais/chap2/ch2s4.htm>

<http://www.crtc.gc.ca/frn/publications/reports/drama/drama2.htm>

<http://www.worldbank.org/wbi/righttotelloverviewfr.html>

Et d'autres...

IN (anglais)

<http://www.pch.gc.ca/progs/ac-ca/progs/ri-bpi/pubs/juneau/anglais/chap2/ch2s4.htm>

<http://www.crtc.gc.ca/eng/publications/reports/drama/drama2.htm>

<http://www.crim.fr/corpus/UK/righttotellOverview.pdf>

Le dessous des cartes (EXEMPLE 2 ^{2/3})

Etapes suivies pour le traitement

• **Traitement des données :**

- Recherche des pages bilingues pertinentes
- Enregistrement des pages en local (wget)
- Mise au format texte des formats .html et .pdf (outils linux)
- Etiquetage des pages françaises (CORDIAL)
- Choix des patrons morpho-syntaxiques pertinents en français
- Extraction des syntagmes conformes aux patrons (PERL)
- Tri des syntagmes obtenus
- Projection des syntagmes sur corpus français
- Alignement fruste avec corpus anglais (PERL)
- Choix des entrées pertinentes et sélection des équivalents en langue-cible

Le dessous des cartes (EXEMPLE 2 ^{3/3})

Résultats

- *OUT*:
 - une proposition de lexique
 - http://www.crim.fr/lexique_media.html
 - une comparaison avec l'existant
 - (c'est pas fait !)

Le dessous des cartes (EXEMPLE 3)

- Le projet CreeATLAS

- <http://netx.u-paris10.fr/defi/CreeATLAS/>



Le projet

- **Parcours**

Une année de projets...

- Au semestre 1
 - Projet multilingue : « *la vie des mots sur le web* »
- Au semestre 2
 - Projet « *Les Boîtes à outils* »
 - Étiquetage, filtrage sur les Fils de Presse
- Fin du semestre 2
 - LE VRAI PROJET
 - Cf projets en ligne des années passées



Le projet

- Je laisse des traces...

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 *pluriTAL*)
 - 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 1 blog (1/3)

<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.

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
sfweb.no-ip.org

ATONET
wiki (TAL-Lexicométrie)

Recherche :

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)

un journal de bord sur 1 blog (2/3)

Une plateforme d'édition en ligne

WordPress

You're lookin' swell, Dolly

Write

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

Logout



Title

← Titre du billet

Connexion sur un compte prédéfini



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

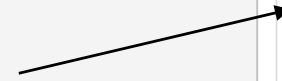
On rédige le billet ici...



Quelques outils d'édition...

On classe le billet...

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



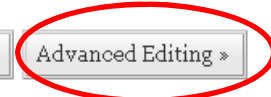
On publie ici : en ligne ou un brouillon

PingBack the URIs in this post ?

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

Save as Draft Save as Private Publish Advanced Editing »

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



un journal de bord sur le blog pluriTAL (3/3)

Le Blog du MASTER : <http://tal-p3.wordpress.com/>

pluriTAL



15 juin 2006

Projet "Multilingue"

Enregistré dans : [Projet 2005-2006, pr-Bilingue \(JMD\)](#) — tal-p3 @ 4:12

Les travaux réalisés dans le cadre du projet dit "MULTILINGUE" sont en ligne sur le site du [CRIM](#) rubrique "[Travaux des étudiants](#)".

- 1) Un [petit lexique français-anglais économique](#), constitué à partir de documents numériques traduits du français vers l'anglais. Les termes choisis ont été extraits automatiquement à partir de patrons morpho-syntaxiques, puis triés à la main. La traduction anglaise des termes a ensuite été trouvée à partir d'un tableau dans lequel les documents français et anglais ont été alignés.
- 2) Un [extrait de lexique français-estonien de l'environnement](#), (à ouvrir avec Mozilla de préférence), constitué à partir de documents traduits du français vers l'estonien. L'extraction des candidats-termes s'est faite à partir de patrons morpho-syntaxiques, puis le lexique a été constitué à partir d'un alignement fin du français et de l'estonien.
- 3) Un [lexique français-anglais de la Formule 1](#), élaboré grâce à des documents bilingues glanés sur les sites "officiels", puis alignés.

(à suivre)

blog Master pluriTAL

**Filières TAL et
ingénierie linguistique
de Paris III Sorbonne
nouvelle, Paris X
Nanterre, INALCO
(Institut National des
langues et civilisations
orientales)**

Contact :

serge.fleury@univ-paris3.fr

Rentrée 2006-2007

[Voir sur le site pluriTAL](#)

Blogroll

- [WordPress.com](#)

Premières activités sur le blog

- Vous vous connectez sur le blog
- Chacun rédige un petit billet de présentation (nom, formation, diplôme préparé, niveau informatique...)

Identifiant : pluritaluser

Mot de passe : ...

My Account ▾ My Dashboard New Post

"pluriTAL" Blog Info ▾ →

pluriTAL

Création d'un billet



Vous ne touchez pas au paramétrage du blog

15 juin 2006

Projet "Multilingue"

Enregistré dans : [Projet 2005-2006](#), [pr-Bilingue \(JMD\)](#) — tal-p3 @ 4:12 [Modifier](#)

Les travaux réalisés dans le cadre du projet dit "MULTILINGUE" sont en ligne sur le site du [CRIM](#) rubrique "[Travaux des étudiants](#)".

1) Un [petit lexique français-anglais économique](#), constitué à partir de documents numériques traduits du français vers l'anglais. Les termes choisis ont été extraits automatiquement à partir de patrons morpho-syntaxiques, puis triés à la main. La traduction anglaise des termes a ensuite été trouvée à partir d'un tableau dans lequel les documents français et anglais ont été alignés.

2) Un [extrait de lexique français-estonien de l'environnement](#), (à ouvrir avec Mozilla de préférence), constitué à partir de documents traduits du français vers l'estonien. L'extraction des candidats-termes s'est faite à partir de patrons morpho-syntaxiques, puis le lexique a été constitué à partir d'un alignement fin du français et de l'estonien.

3) Un [lexique français-anglais de la Formule 1](#), élaboré grâce à des documents bilingues glanés sur les sites "officiels", puis alignés.

blog Master pluriTAL

*Filières TAL et
ingénierie linguistique
de Paris III Sorbonne
nouvelle, Paris X
Nanterre, INALCO
(Institut National des
langues et civilisations
orientales)*

Contact :

serge.fleury@univ-paris3.fr

Retournée 2006-2007

Voir sur le site pluriTAL

Blogroll

• [WordPress.com](#)• [WordPress.org](#)

Titre

Début des cours le 12/10/2006

Classement des billets

Article

Editeur HTML

Menu d'édition



Le cours "Programmation et Projet Encadré" redémarre le 12/12/2006.

Bon travail à tous !

Interface d'édition

Chemin: p

Sauvegardé à 2:54:20

Sauvegarde continue

Enregistrer

Publier

Catégories

Ajouter

Séparez les catégories multiples par des espaces

- Projet 2006-2007
- Blogroll
- Master TAL Recherche
- metaBlog
- plurITAL
- Projet 2005-2006
 - pr-Bilingue (JMD)
 - pr-Communautés

Discussion

- Autoriser les commentaires
- Autoriser les Pings

Charger Tout voir Video

File Parcourir...

Titre

Description

Mode de publication

Charger »

Mot de passe de la note +

Post Slug +

État de l'article

- Publié
- Brouillon
- Privé



Le projet

- Une problématique
- Des outils
- Des méthodes
- etc

[sommaire](#)

Définition d'un problème linguistique (pour le projet final)

- Une liste « imposée »
 - La notion de discret vs. continu en anglais selon le type de discours
 - Utilisation des verbes et des substantifs dans des corpus parallèles français-anglais
 - Analyse morphologique du discours scientifique
 - Dégrouper les sens (corpus sur le web ou corpus de presse) :
 - <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/PROJETSM1/ProjetDegrouperSensWebLeMonde.pdf>(présentation)
 - Nuages de mots dans le Fils RSS :
 - Présentation du projet 2005-2006 ici : <http://tal.univ-paris3.fr/filspresse/>
 - <http://tal.univ-paris3.fr/filspresse/projet-fils-de-presse.pdf>
- Ou autre chose...
 - À valider par nous...

Collecte des données

- Des outils

- wget, httrack, lynx, LWP etc.

- Lectures des manuels

- Wget :

- <http://www.gnu.org/software/wget/manual/>

- Httrack : <http://www.httrack.com/html/index.html>

- Lynx :

- <http://www.ldh.org/Dossiers/Manuels/lynx.html>

- Automatisation

- cron

Formatage des données

- Hétérogénéité des formats
 - HTML, PDF, PS, etc.
 - Redéfinition de la récolte ?
 - Qualité des données selon les formats ?
- Normalisations des données
 - HTML2TXT, PDF2TXT, PS2TXT, etc.
 - Des scripts : PERL

Analyse simple des données

- Compter
 - wc, uniq
 - Lecture manuel
- Filtrer
 - egrep
 - Lecture manuel
- Trier
 - sort
 - Lecture manuel
- Ne pas oublier « les expressions régulières »
 - Lecture manuel

Analyse avancée des données

- Morphologie
- Etiquetage syntaxique
- Extraction de patrons
- Etc.

Organisation des données

- Données collectées
- Données formatées
- Données produites

- Structurer les données « manipulées » pour les retrouver, les traiter, les présenter...
 - Fichier, répertoire, systèmes de fichier, sauvegarde

Vers une présentation des données...

- Passage des résultats à une présentation linguistique
- Affichage des résultats
 - Structures les tâches réalisées et les résultats produits
 - Rédiger des rapports d 'analyse
- Synthèse finale

Des pauses d'apprentissage...

- Des rappels
 - Systèmes de fichier
 - Les expressions régulières
 - Langage de script : Perl
 - Commandes unix
 - Etc.

- Bon travail.....



Les outils

- Préambule : cygwin

Démarrer avec Cygwin

- Au LABOC :
 - Une version de *Cygwin* « brute », *i.e.* uniquement la fenêtre de commandes !
 - Démarrage :
 - *Menu Démarrer >> Cygwin >> Cygwin Bash Shell*
 - une version de Cygwin (dite « *cygwin-X* ») « avec interface graphique » est installée
 - Démarrage :
 - *Menu Démarrer >> ... >> « Start-X Server »*
 - On peut ensuite lancer les applications disponibles (*cf* menu démarrer >> *Cygwin-X*)

Préambule

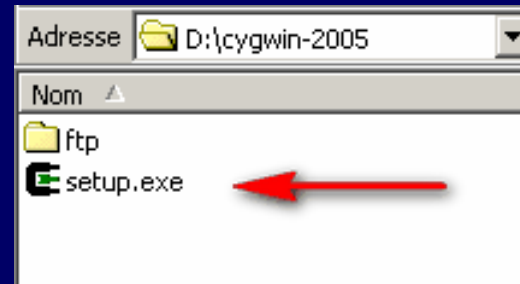
- **Cygwin** n'est ni plus ni moins que le portage sous windows du compilateur GNU **GCC** que l'on retrouve sous Linux et d'autres UNIX. Il permet d'écrire du code en C et C++, code qui pourra être utilisé sur une autre machine Windows ne disposant pas de **Cygwin**.
- Mais **Cygwin** c'est bien plus que cela, c'est un véritable OS dans l'OS, il copie complètement UNIX avec toutes ces commandes standards et son environnement, c'est donc une excellente alternative pour découvrir UNIX.
- **Cygwin** est un produit en plein développement, ne vous étonnez donc pas s'il peut planter de temps à autre.
- Pour la petite histoire **cygwin** est un produit de la société **cygnus**, qui entre autres réalise le célèbre compilateur GNU **GCC**, et qui a été acheté récemment par **Redhat**, d'où le format de l'URL du site officiel <http://sources.redhat.com/cygwin>.
- La licence est la suivante, si vous avez acheté une licence avec **Cygnus**, vous avez le droit de créer des programmes mais ne peut diffuser le code, par contre si vous n'avez pas acheté de licence, vous tombez dans le cadre de la licence GPL, c'est à dire que vous êtes obligé pour tout programme que vous créez et fournissez de fournir aussi le code associé.

Installation

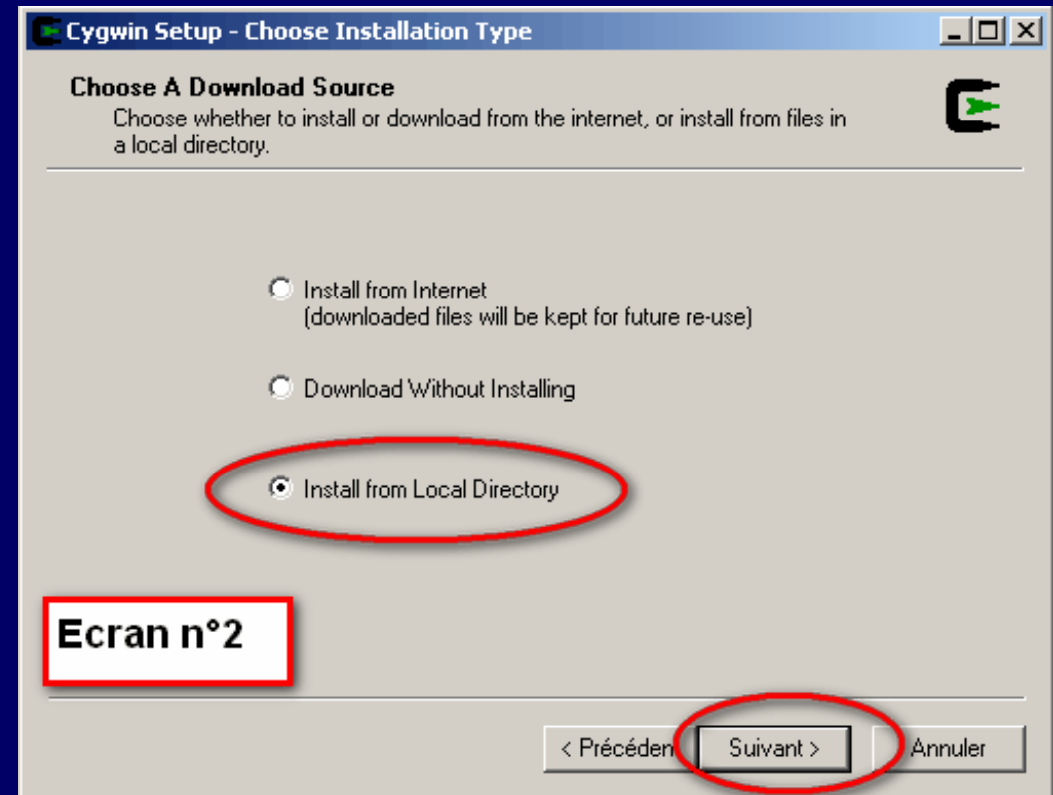
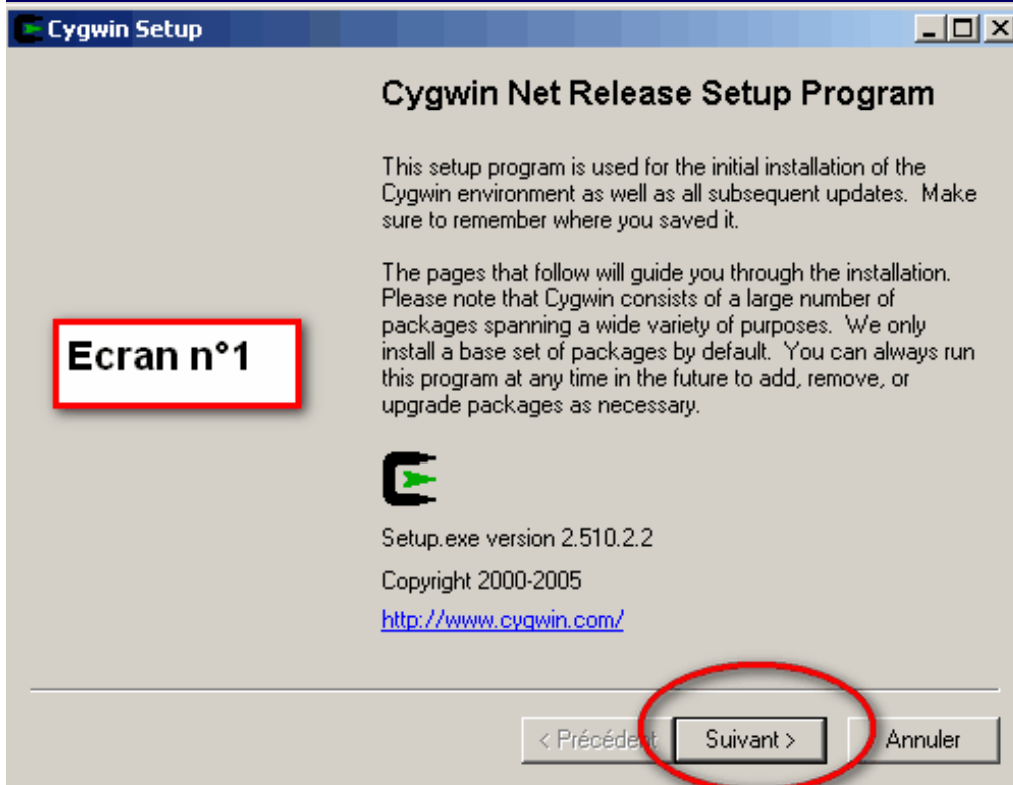
- Vous avez la possibilité de récupérer les packages qui vous intéressent directement sur les sites miroirs de **cygwin** et de vous débrouiller pour les installer, ou alors de télécharger un petit exécutable qui va vous guider dans la récupération des packages et leur installation.
- On récupérera **setup.exe** à l'URL <http://sourceware.org/cygwin/> .
- A l'exécution de ce programme, on passe d'abord par une fenêtre d'information, cliquez sur **Next**. Vous pouvez alors choisir de télécharger dans un premier temps les packages (**Download from Internet**), de télécharger et d'installer dans la foulée (**Install from Internet**), d'installer les packages qui ont été téléchargés préalablement et qui sont contenus dans un répertoire local (**Install from Local directory**)... on se laisse guider (pas de difficultés)



- Je tiens à votre disposition un CD avec l'ensemble des ressources telles qu'elles sont installées au LABO...



Un exemple d'installation de Cygwin



Cygwin Setup - Choose Installation Directory

Select Root Install Directory
Select the directory where you want to install Cygwin. Also choose a few installation parameters.

Root Directory
 Browse...

Install For

- All Users (RECOMMENDED)**
Cygwin will be available to all users of the system. **NOTE:** This is required if you wish to run services like sshd, etc.
- Just Me
Cygwin will only be available to the current user. Only select this if you lack Admin. privileges or you have specific needs.

Default Text File Type

- Unix / binary (RECOMMENDED)**
No line translation done; all files opened in binary mode. Files on disk will have LF line endings.
- DOS / text
Line endings will be translated from unix (LF) to DOS (CR-LF) on write and vice versa on read.
[Read more about file modes...](#)

Ecran n°3

< Précédent **Suivant >** Annuler

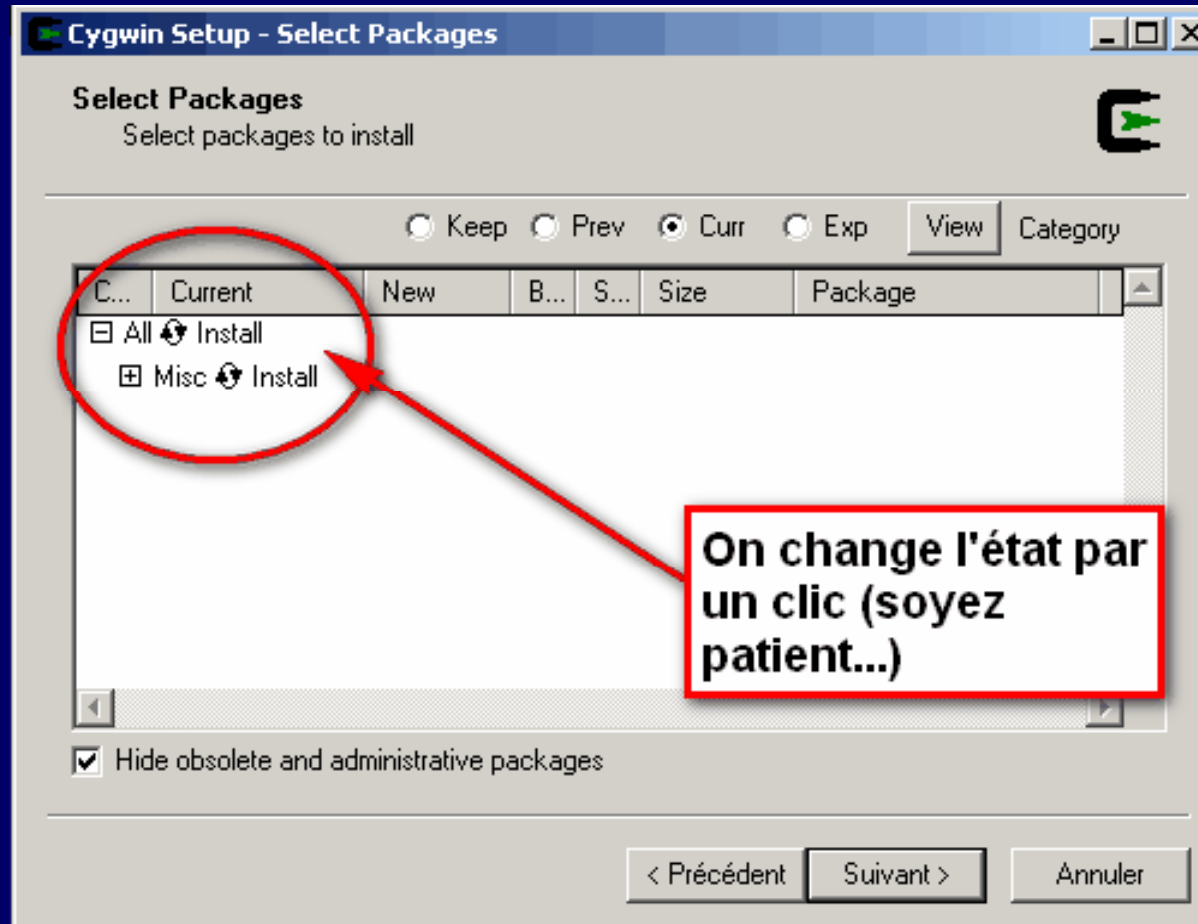
Cygwin Setup - Select Local Package Directory

Select Local Package Directory
Select a directory where you want Setup to store the installation files it downloads. The directory will be created if it does not already exist.

Local Package Directory
 Browse...

Ecran n°4

< Précédent **Suivant >** Annuler



Select Packages

Select the packages you want setup to install.

Keep
 Prev
 Curr
 Exp
 View Category

Category	Curr...	New	Bi...	Sr...	Package
+ All		Default			
+ Admin		Default			
+ Archive		Default			
+ Base		Default			
+ Database		Default			
+ Devel		Default			
+ Doc		Default			
+ Editors		Default			
+ Games		Default			
+ Graphics		Default			

< Précédent Suivant > Annuler



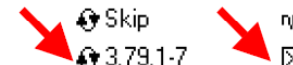
Select Packages

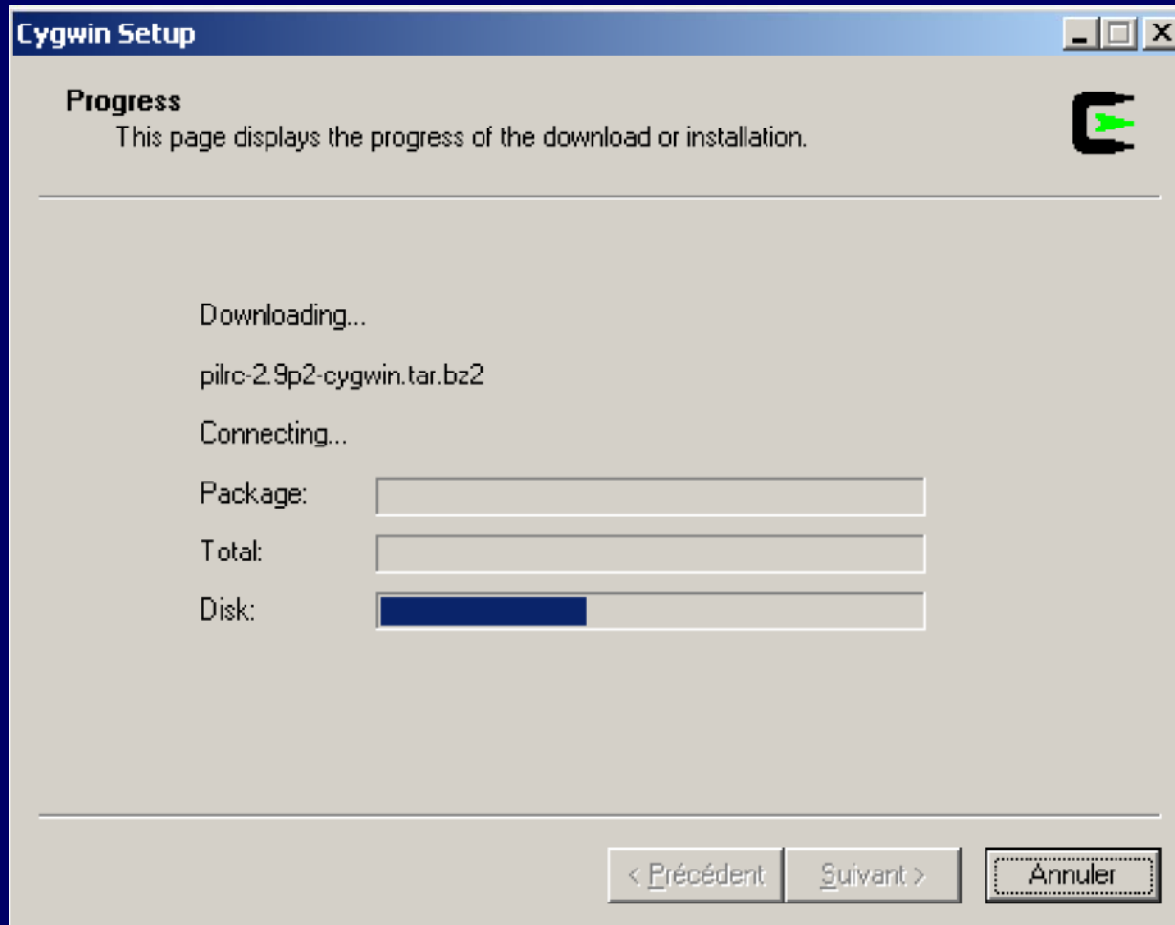
Select the packages you want setup to install.

Keep
 Prev
 Curr
 Exp
 View Category

Category	Curr...	New	Bi...	Sr...	Package
		<input checked="" type="checkbox"/> Skip	n/a	n/a	libncurses-devel: Libraries
		<input checked="" type="checkbox"/> Skip	n/a	n/a	libtool: Wrapper scripts for
		<input checked="" type="checkbox"/> Skip	n/a	n/a	libtool-devel: A shared libr
		<input checked="" type="checkbox"/> Skip	n/a	n/a	libtool-stable: A shared libr
		<input checked="" type="checkbox"/> Skip	n/a	n/a	libxml2: Libxml is the XML
		<input checked="" type="checkbox"/> Skip	n/a	n/a	libxslt: Libxslt is the XSLT
		<input checked="" type="checkbox"/> 3.79.1-7	<input checked="" type="checkbox"/>	<input type="checkbox"/>	make: The GNU version of
		<input checked="" type="checkbox"/> Skip	n/a	n/a	mingw-runtime: MinGW R
		<input checked="" type="checkbox"/> Skip	n/a	n/a	mktemp: Allows safe temp
		<input checked="" type="checkbox"/> Skip	n/a	n/a	nasm: The Netwide Assem

< Précédent Suivant > Annuler





Ressources:

<http://cygwin.com/cygwin-ug-net/setup-net.html>: Setting up Cygwin

<http://cygwin.com/cygwin-ug-net/using.html>: Using Cygwin

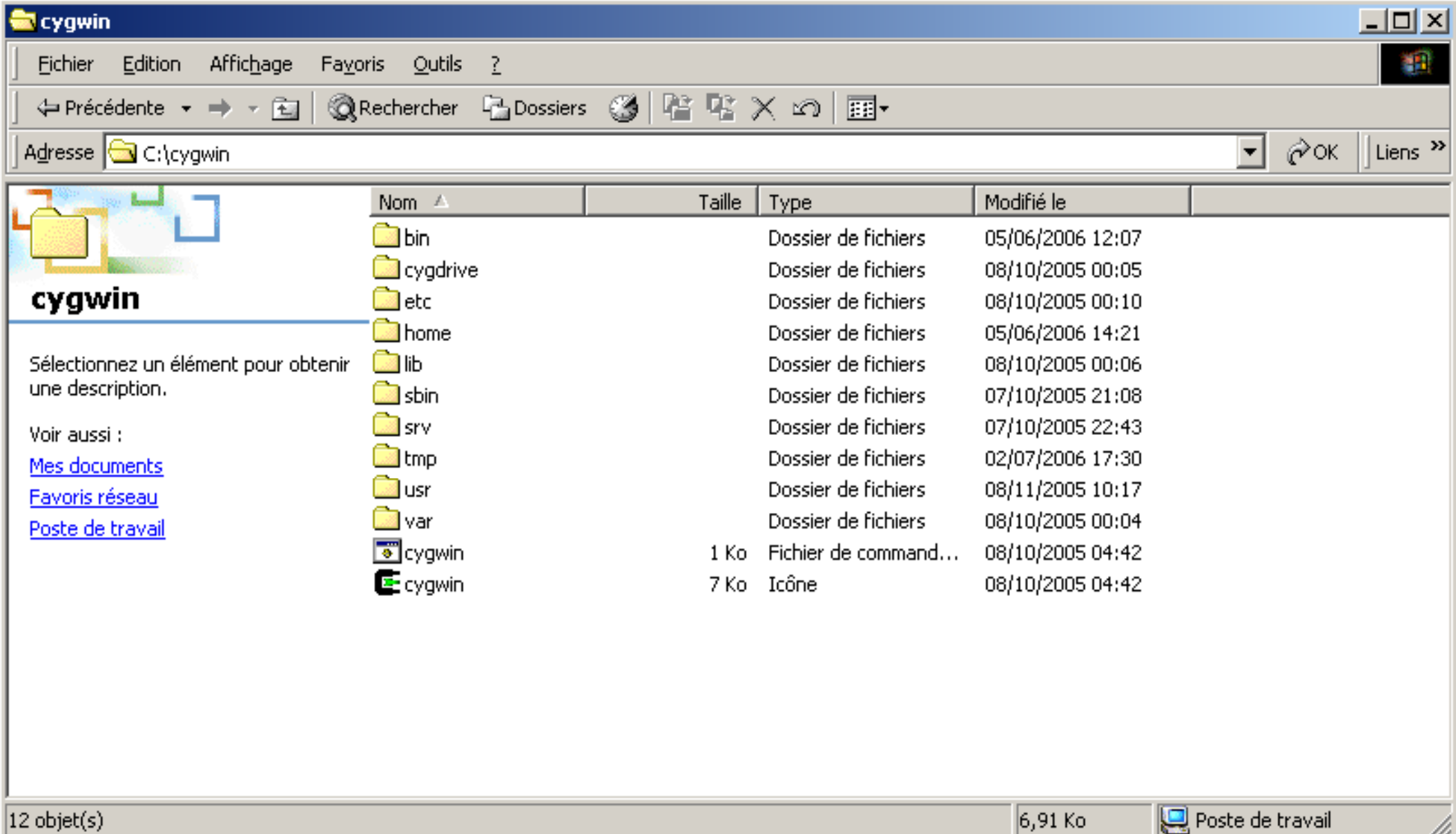
Une fois l'installation finie



Arborescence cygwin

- Les sous répertoires suivants sont créés dans le répertoire d'installation: - **bin**, contenant les exécutables, on retrouve beaucoup de commandes standards UNIX, en voici un brel panel : **awk, bash, bison, bunzip2, chmod, date, dd, find, grep, kill, m4, make, gcc, man, mount, rcp, su, telnet, touch, vi, whoami**, et j'en passe et des meilleurs...
 - **etc**, contenant les fichiers de configuration.
 - **home**, contenant les répertoires utilisateurs
 - **lib**, contenant toutes les bibliothèques diverses y compris celle de développement
 - **tmp**, répertoire temporaire
 - **usr** qui contient lui-même
 - **bin**, répertoire d'exécutable (vide par défaut)
 - **doc**, pour documentation
 - **include**, contenant les headers nécessaires pour développer
 - **info**, (?)
 - **lib**, autre répertoire de bibliothèque (vide par défaut)
 - **libexec**, répertoire d'exécutable
 - **local** contenant d'autres sous répertoires (bin, doc, etc, include et lib) pour certaines applis bien spécifiques
 - **man**, contient les manuels des divers applis
 - **sbin**, exécutables pour l'administrateur
 - **share**, contient les ressources des différents exécutables
 - **ssl**, pour rajouter une couche sécurisée à la couche transport de TCP
 - **tmp**, autre répertoire temporaire
 - **var**, contient les variables temporaires et autres fichiers de logs.
- Vous avez vite reconnu une arborescence type UNIX

Arborescence Cygwin



The screenshot shows a Windows Explorer window titled "cygwin" with the address bar set to "C:\cygwin". The main pane displays a list of files and folders. The left pane shows a tree view with "cygwin" selected. The status bar at the bottom indicates "12 objet(s)", "6,91 Ko", and "Poste de travail".

Nom	Taille	Type	Modifié le
bin		Dossier de fichiers	05/06/2006 12:07
cygdrive		Dossier de fichiers	08/10/2005 00:05
etc		Dossier de fichiers	08/10/2005 00:10
home		Dossier de fichiers	05/06/2006 14:21
lib		Dossier de fichiers	08/10/2005 00:06
sbin		Dossier de fichiers	07/10/2005 21:08
srv		Dossier de fichiers	07/10/2005 22:43
tmp		Dossier de fichiers	02/07/2006 17:30
usr		Dossier de fichiers	08/11/2005 10:17
var		Dossier de fichiers	08/10/2005 00:04
cygwin	1 Ko	Fichier de command...	08/10/2005 04:42
cygwin	7 Ko	Icône	08/10/2005 04:42

Lancement de Cygwin (1)

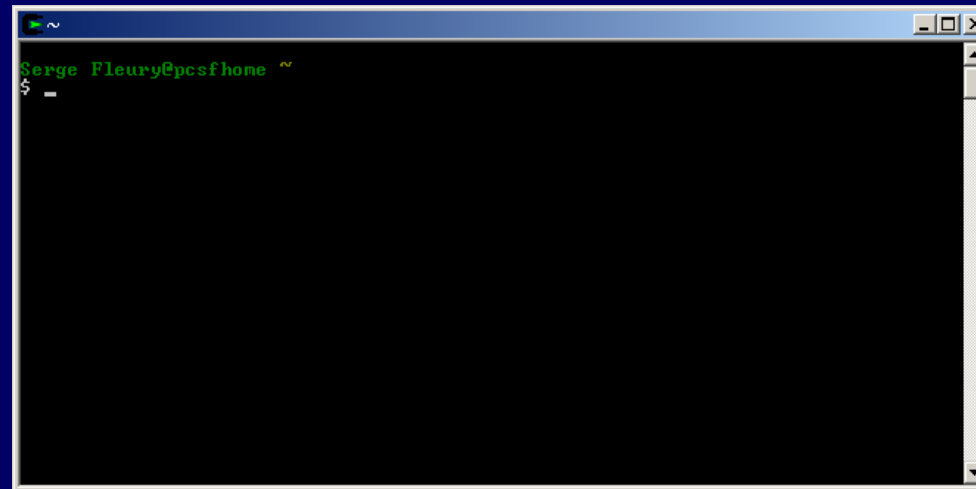
- On lance cygwin à partir de l'icône sur le bureau ou du menu Démarrer
 - Programmes->Cygnus Solutions->Cygwin Bash Shell
- une fenêtre type DOS apparaît qui est en fait un *shell* comme on peut en trouver sous UNIX, c'est comme son nom l'indique l'implémentation sous windows du *bash shell* bien connu sous Linux.
- Ceux qui sont familiers à UNIX n'auront aucun mal à se débrouiller avec ce *shell*, les autres devront commencer peut être d'abord à apprendre les commandes de base d'UNIX.
- Vous aurez vite constaté que comme sous Unix vous devez pour vous déplacer dans l'arborescence taper / comme sous UNIX au lieu de \. Vous retrouvez en vous baladant dans l'arborescence, celle qui a été présentée dans le paragraphe précédent.

Lancement de Cygwin (2)

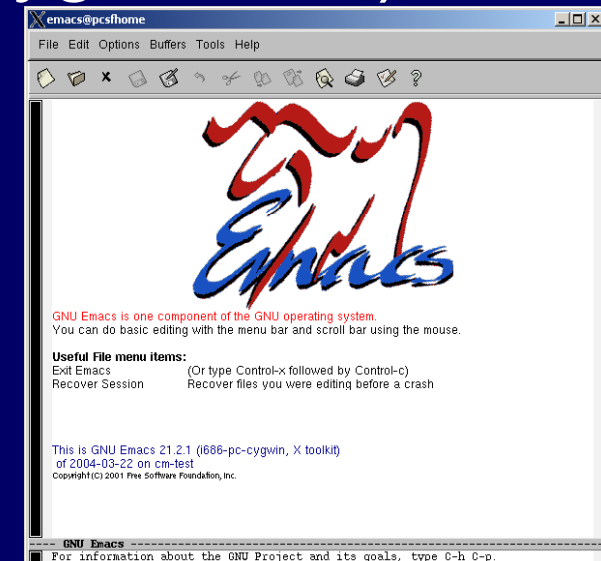
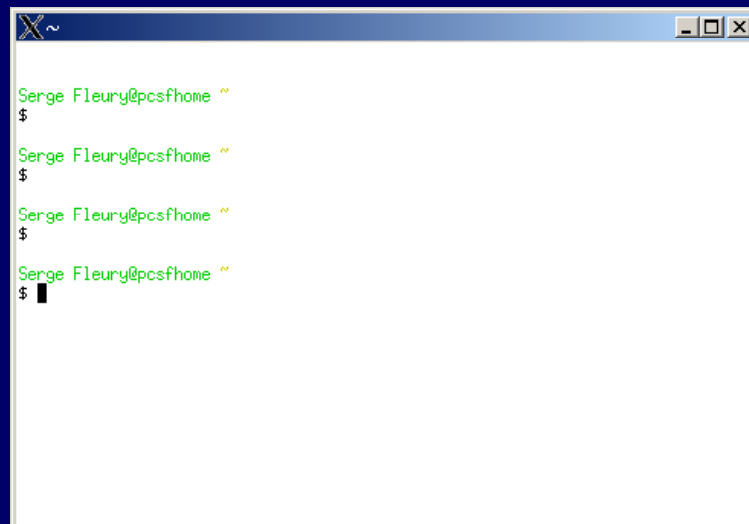
- Une version de *Cygwin* « brute », *i.e.* uniquement la fenêtre de commandes !
 - Démarrage :
 - *Menu Démarrer >> Cygwin >> Cygwin Bash Shell*
- une version de Cygwin (dite « cygwin-X ») « avec interface graphique » est installée
 - Démarrage :
 - *Menu Démarrer >> ... >> « Start-X Server »*
 - On peut ensuite lancer les applications disponibles (*cf* menu démarrer >> Cygwin-X)

Lancement de Cygwin (3)

- Une version de *Cygwin* « brute »

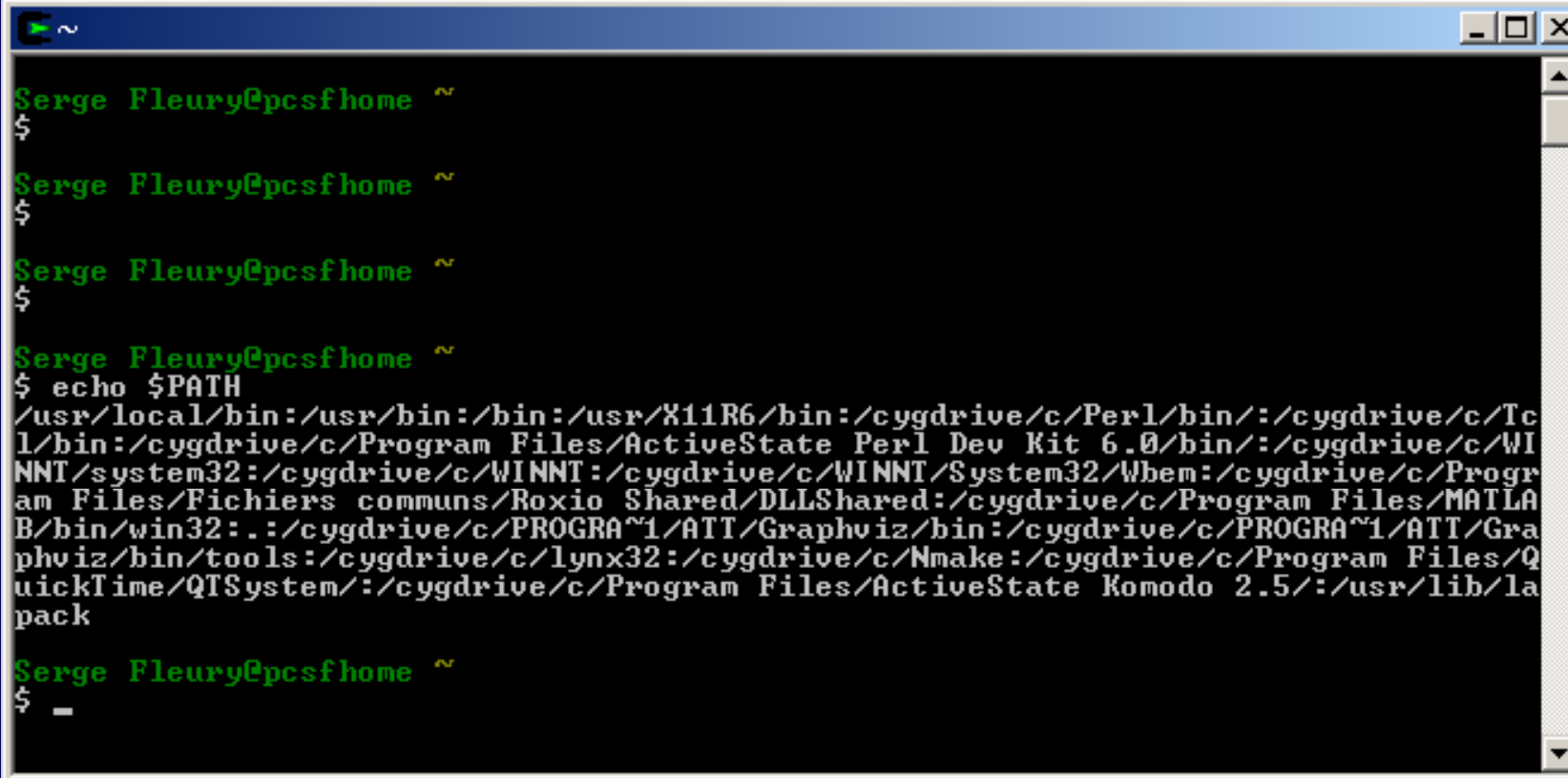


- une version de Cygwin (dite « cygwin-X »)



Remarque

Comment *cygwin* a-t-il accès aux différents disques du PC ? Regardons la variable `PATH` sous *cygwin* :



```
Serge Fleury@pcsfhome ~
$
Serge Fleury@pcsfhome ~
$
Serge Fleury@pcsfhome ~
$
Serge Fleury@pcsfhome ~
$ echo $PATH
/usr/local/bin:/usr/bin:/bin:/usr/X11R6/bin:/cygdrive/c/Perl/bin:/cygdrive/c/Tcl/bin:/cygdrive/c/Program Files/ActiveState Perl Dev Kit 6.0/bin:/cygdrive/c/WINNT/system32:/cygdrive/c/WINNT:/cygdrive/c/WINNT/System32/Wbem:/cygdrive/c/Program Files/Fichiers communs/Roxio Shared/DLLShared:/cygdrive/c/Program Files/MATLAB/bin/win32:./:/cygdrive/c/PROGRA~1/ATT/Graphviz/bin:/cygdrive/c/PROGRA~1/ATT/Graphviz/bin/tools:/cygdrive/c/lynx32:/cygdrive/c/Nmake:/cygdrive/c/Program Files/QuickTime/QTSystem:/cygdrive/c/Program Files/ActiveState Komodo 2.5:/usr/lib/lpack
Serge Fleury@pcsfhome ~
$ _
```

Pour avoir accès au disque D, on écrit `/cygdrive/d/`. Essayez : `ls -l /cygdrive/d/`

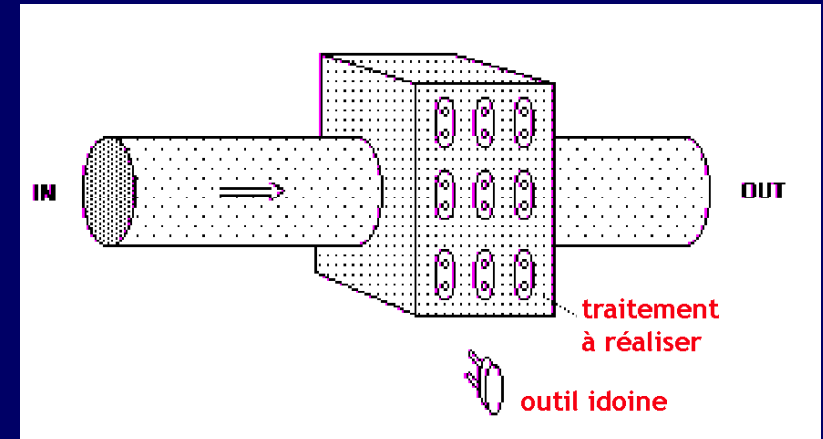


Projet n°1

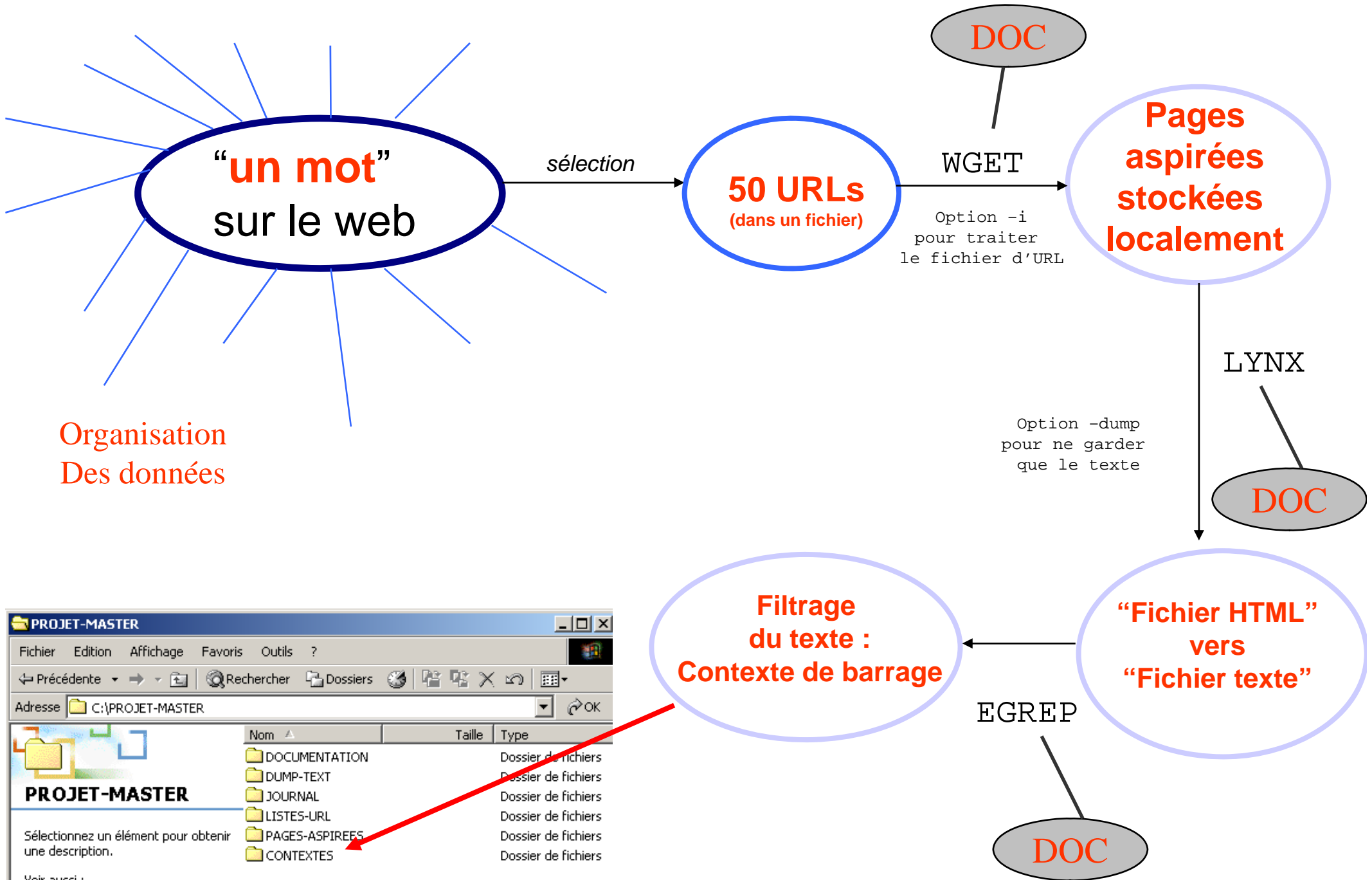
La vie « multilingue »
des mots sur le web

[sommaire](#)

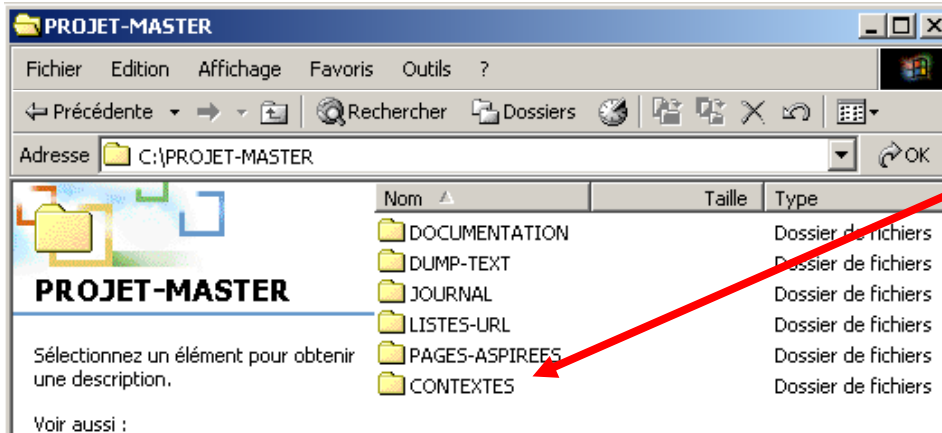
Esquisse de parcours pour le projet n°1 à construire



- Objectif : construire une chaîne complète de traitements
 - Définition du problème linguistique visé
 - « la vie multilingue des mots sur le web »
 - Choix des données à traiter
 - Un choix raisonné d'URL contenant un mot choisi (dans plusieurs langues)
 - Collecte des données
 - Formatage des données
 - Utiliser les outils « proposés », apprendre à les utiliser...
 - Wget, lynx, egrep (commandes unix), scripts bash...
 - Produire des résultats et les présenter
 - Un site web contenant : rapport d'activités + résultats



Organisation
Des données



Exemples de projet les années passées...

- 2007-2008

- <http://tal.univ-paris3.fr/plurital/travaux-2007-2008/tr0708-masterproj-sem1.htm>

- 2006-2007

- <http://tal.univ-paris3.fr/plurital/travaux-2006-2007/tr0607-masterproj-sem1.htm>

- 2005-2006

- <http://tal.univ-paris3.fr/plurital/travaux-2005-2006/masterproj/tr0506-masterproj-sem1.htm>

Examen d'un mot sur le web :

- le mot « barrage » (ou un mot de votre choix) et ses traductions possibles...
 - Le mot « barrage » dans le TLFi (cf slide suivant) :
 - <http://www.cnrtl.fr/lexicographie/barrage?>
 - Co-occurents « barrage » :
 - lien slide :
 - <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/PROJETSM1/LMPCoocCoocBarrage220PremiersIDF.pdf>
 - Lectures :
 - « Cartographie lexicale pour la recherche d'information », Jean Véronis, 2003
 - <http://www.up.univ-mrs.fr/veronis/pdf/2003-taln.pdf>
 - Voir aussi HYPERLEX : <http://www.up.univ-mrs.fr/veronis/demos/index.html?http://www.up.univ-mrs.fr/veronis/demos/hyperlex.html>
 - « Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales », Olivier Ferret, 2004
 - <http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Ferret.pdf>

Barrage dans le TLFi



Laboratoire d'Analyse et de Traitement Informatique de la Langue Française
Trésor de la Langue Française informatisé (version simplifiée)

Nouvelle recherche



Signification des couleurs

Mot recherché

Expressions ou locutions

Définitions

BARRAGE, subst. masc.

A. — Action de barrer.

— Spécialement

1. **TECHN. MILIT.** Tir de barrage. Tir d'artillerie destiné à barrer le passage à l'ennemi.

2. **PSYCHOL.** [Correspond à barrer II] *Test de barrage*. Test de l'attention qui consiste à barrer d'un trait certains signes géométriques simples, mélangés à d'autres signes presque semblables (d'apr. *Psychol.* 1969).

3. **SP.** *Match de barrage*. Match servant à départager deux concurrents ou deux équipes qui, au cours d'une compétition générale (championnat, coupe, etc.), sont à égalité de points.

B. — *P. méton.* **Barrière, obstacle qui ferme un passage.** *Barrage de police; forcer un barrage :*

• 1. ... vers le milieu de cette ruelle il se heurta à un **obstacle**. Il étendit les mains. C'était une charrette renversée; son pied reconnu des flaques d'eau, des fondrières, des pavés épars et amoncelés. Il y avait là une **barricade** ébauchée et abandonnée. Il escalada les pavés et se trouva de l'autre côté du **barrage**.

HUGO, *Les Misérables*, t. 2, 1862, p. 351.

• 2. ... il est interdit : 1 De placer dans les cours d'eau aucun **barrage**, appareil ou établissement quelconque de pêcheries ayant pour objet d'empêcher entièrement le passage du poisson. (Loi du 15 avril 1829, art. 24).

Code de la pêche fluviale, 1875, p. 91.

• 3. De quelque côté qu'on tente de tourner ce lit, un obstacle vous rejette, ou Christine, ou la bonne sœur, ou l'**empilement** des oreillers, ou le **barrage** des rideaux; ...

MALÈGUE, *Augustin*, t. 2, 1933, p. 363.

— *P. métaph.* :

• 4. Le christianisme, par exemple, n'est plus aujourd'hui qu'un **barrage**, une *pyramide en travers* du chemin, une *montagne* de pierres qui *entrave* les constructions nouvelles.

RENAN, *L'Avenir de la sc.*, 1890, p. 383.

• 5. André avait la faculté de ne souffrir qu'en surface; son inattention volontaire opposait à sa déception un **barrage** étanche, l'empêchait d'atteindre la conscience.

R. MARTIN DU GARD, *Devenir*, 1909, p. 84.

• 6. La censure est un **barrage** psychique qui aboutit à un compromis, exige le remaniement, le déguisement des désirs refoulés...

RICŒUR, *Philos. de la volonté*, 1949, p. 359.

1. **DR. FÉOD.** **Barrière qu'on ne pouvait franchir qu'en payant un droit de péage.** *Droit de barrage :*

• 7. Quelques châteaux situés sur le bord d'une rivière levaient un **impôt** sur la navigation, au moyen d'un **barrage** ou *estacade* qui ne laissait un *passage* qu'assez près des remparts pour que les bateaux ne pussent se soustraire au paiement du droit fixé. Il y avait, par exemple, un **barrage** sur la Seine auprès du Château-Gaillard.

MÉRIMÉ, *Ét. sur les arts au Moy. Âge*, 1870, p. 233.

Rem. La plupart des dict. du XIX^e s. (Ac. jusqu'à 1878) ainsi que QUILLET 1965 enregistrent un subst. masc. *barrager*, vx, „Celui qui était chargé de percevoir le droit de barrage.“

2. **TRAV. PUBL.** **Ouvrage construit sur un cours d'eau, soit pour dériver ou régulariser le cours, soit pour en utiliser la retenue comme source d'énergie ou comme réserve pour l'irrigation :**

• 8. Il était possible de corriger le défaut de pente et l'irrégularité de débit par la construction de **barrages** -réservoirs, énormes *ouvrages barrant* des vallées entières : on est parvenu ainsi, à grand-peine et à grands frais, à emmagasiner des millions de mètres cubes.

J. BRUNHES, *La Géogr. hum.*, 1942, p. 36.

• 9. Il reste à parler des **barrages** qui ne créent pas à proprement parler des retenues, mais qui servent principalement à *dériver* les eaux dans le canal d'amenée. On les appelle, de préférence, « **barrages de dérivation** ».

G. THALLER, *La Houille blanche*, 1952, p. 48.

SYNT. *Barrage d'accumulation, de prise, de régulation, de retenue; barrage fixe, mobile, provisoire, artificiel.*

— *P. anal.* :

• 10. On entendait aussi le bruit assez accentué d'une chute d'eau, qui indiquait, à quelques cents pas en amont, la présence d'un **barrage naturel**.

VERNE, *L'Île mystérieuse*, 1874, p. 237.

Prononc. : [ba.ʁaːʒ] ou [ba-].

Étymol. ET HIST. — 1. 1130-60 « ce qui ferme un passage, barrière, porte » (*Couronnement Louis*, éd. E. Langlois, 436 dans T.-L. : Si come il vindrent, hurtent par lor otrage A la grant porte, qui n'esteit mie basse : Seignor baron, dist l'apostoles sages, Ici endroit guarderez cest barnage [lire **barrage**?]), attest. isolée; 1835, Ac.; **a**) 1363 « droit de passage » (*Ordinat. reg. Franc.*, t. 4, p. 729 dans DU CANGE : oudit lieu de la chancie ledit **Barraige** soit cueillis, levés et exploités sur les passans); **b**) p. anal. 1915 *tir de barrage* « tir destiné à barrer le passage à une troupe » (P. BOURGET, *Le Sens de la mort*, p. 195); 2. 1669 « barrière qui ferme une rivière » (*Ord.*, tit. XXVII, Art. 42 d'apr. VAUVR. *Pêches* qui résume le contenu de cette ordonnance et précise „une décision du ministre des finances du 5 janvier 1815 — au *Recueil des Réglements forestiers* — autorise les préfets à ordonner la suppression des **barrages** établis sur la Loire“); 1842 (HUGO, *Le Rhin*, p. 433 : Et là, debout sur ce magnifique **barrage** naturel qui clôt la Mer Egée, fermant aux Turcs la sortie de l'Archipel).

Dér. de *barre**; suff. -age*.

STAT. — **Fréq. abs. littér.** : 404. **Fréq. rel. littér.** : XIX^e s. : a) 64, b) 516; XX^e s. : a) 600, b) 1 025.

BGG. — **DUB.** Dér. 1962, p. 30.

Accès au TLFi

- <http://atilf.atilf.fr/tlf.htm>
- Lien direct à partir :
 - Du site pluriTAL
 - <http://tal.univ-paris3.fr/plurital/>
 - Du site TAL-ILPGA :
 - <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/index.htm>

Premières activités... (1)

• Rechercher le mot sur le web

– Moteur de recherche sur le web

- Google : <http://www.google.fr> (moteur généraliste)
- Yahoo : <http://fr.yahoo.com/> (moteur généraliste)
- Mozbot : <http://mozbot.fr/> (moteur généraliste francophone, cf Google)
- Google(Blogs) : <http://blogsearch.google.com/blogsearch> (moteur de blogs)
- Google scholar : <http://scholar.google.com/> (moteur « académique »)
- Google print : <http://print.google.com> (moteur de livres)
- CiteSeer : <http://citeseer.ist.psu.edu/> (moteur de littérature scientifique)
- Etc.

Premières activités... (2)

- Rechercher le mot (2)
 - Moteur sur Corpus du Monde électronique
 - Depuis avril 2003, indexation locale, 100.000.000 formes
 - Lieu <http://sfmac.no-ip.com/corpusLeMonde/>
 - Login :
 - Mot de passe :
 - Item « Search » dans le menu de gauche
 - Les pages indexées sont les pages HTML visibles individuellement derrière l'item « corpus HTML » du même menu de gauche.
 - Accès à chaque version quotidienne du journal au format HTML

Premières activités... (3)

- Rechercher le mot (3)
 - Fils de news RSS
 - Fichiers RSS (ATOM) : Description structurée de ressources mises à jour sur un site
 - Par exemple, Fils de News sur le site du Monde
 - <http://www.lemonde.fr/rss/sequence/0,2-3208,1-0,0.xml> (A la une)
 - <http://www.lemonde.fr/rss/sequence/0,2-3210,1-0,0.xml> (International)
 - <http://www.lemonde.fr/rss/sequence/0,2-3224,1-0,0.xml> (France)
 - Pour Libération
 - <http://www.liberation.fr/rss.php>
 - Pour Le Figaro (46 fils)
 - http://www.lefigaro.fr/rss/figaro_une.xml (la Une)
 - Annuaire de Fils RSS : <http://www.tout-en-ligne.com/>
 - **Pubsub**
 - Monitor over over 16 million blog postings and news feeds in real time...
 - Méta-moteur de recherches permet d'effectuer des recherches à travers des fils RSS depuis des mots clés définis par l'utilisateur du service
 - **Newsnetplus** : <http://www.newsnetplus.com/>
 - est un moteur de recherche permettant de trouver des fils RSS en fonction de vos mots-clés. On peut classer les résultats par date ou par pertinence et il est aussi possible d'obtenir une prévisualisation des résultats. Comme Pubsub ou Technocrati, ce nouveau moteur génère un fil RSS à chacune de vos requête que vous pouvez enregistrer dans votre agrégateur afin de surveiller les nouveautés.
 - Annares d'outils pour fils RSS et BLOGS
 - http://inforizon.blogs.com/veille/outils_blogs_et_fils_rss/

Premières activités... (4)

- Essayer de récupérer les pages « pertinentes »
 - Comment faites-vous ?
 - À la main ou non
 - Outils de collectes : `wget...` *cf infra*
 - Difficultés rencontrées
 - Type de fichiers « pêchés » ?
 - L'état du texte ?

Premières activités... (5)

• Pour le projet 2008-2009

- Les recherches menées sur le web doivent conduire *in-fine* à construire un CORPUS MULTILINGUE (2 ou + langues)

– Outils pour un recherche multilingue

- Cf ressources fournies sur la page du cours
 - <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/masterproj.htm>
- **Widepress** : recherche dans 500 journaux en ligne
 - <http://www.widepress.com>
- **Babelplex**, un service qui traduit vos requêtes Google dans des tas de langues différentes et vous permet de les lancer directement.
 - <http://www.babelplex.com/>
- **Wikimindmap** : Outil qui permet de réaliser une carte heuristique à partir d'un mot, d'un terme utilisé dans wikipedia. Très facile d'utilisation et fonctionne pour wikipedia dans plusieurs langues.
 - <http://wikimindmap.org/>

Préparation du rapport

- Rédigez un premier rapport sur les activités réalisées...
 - Le premier élément du rapport final !!!!!

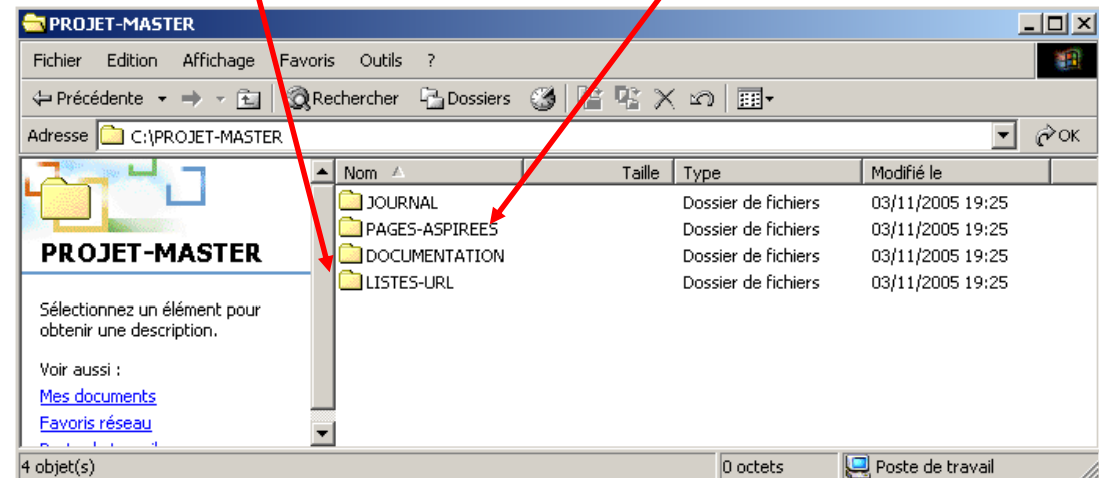
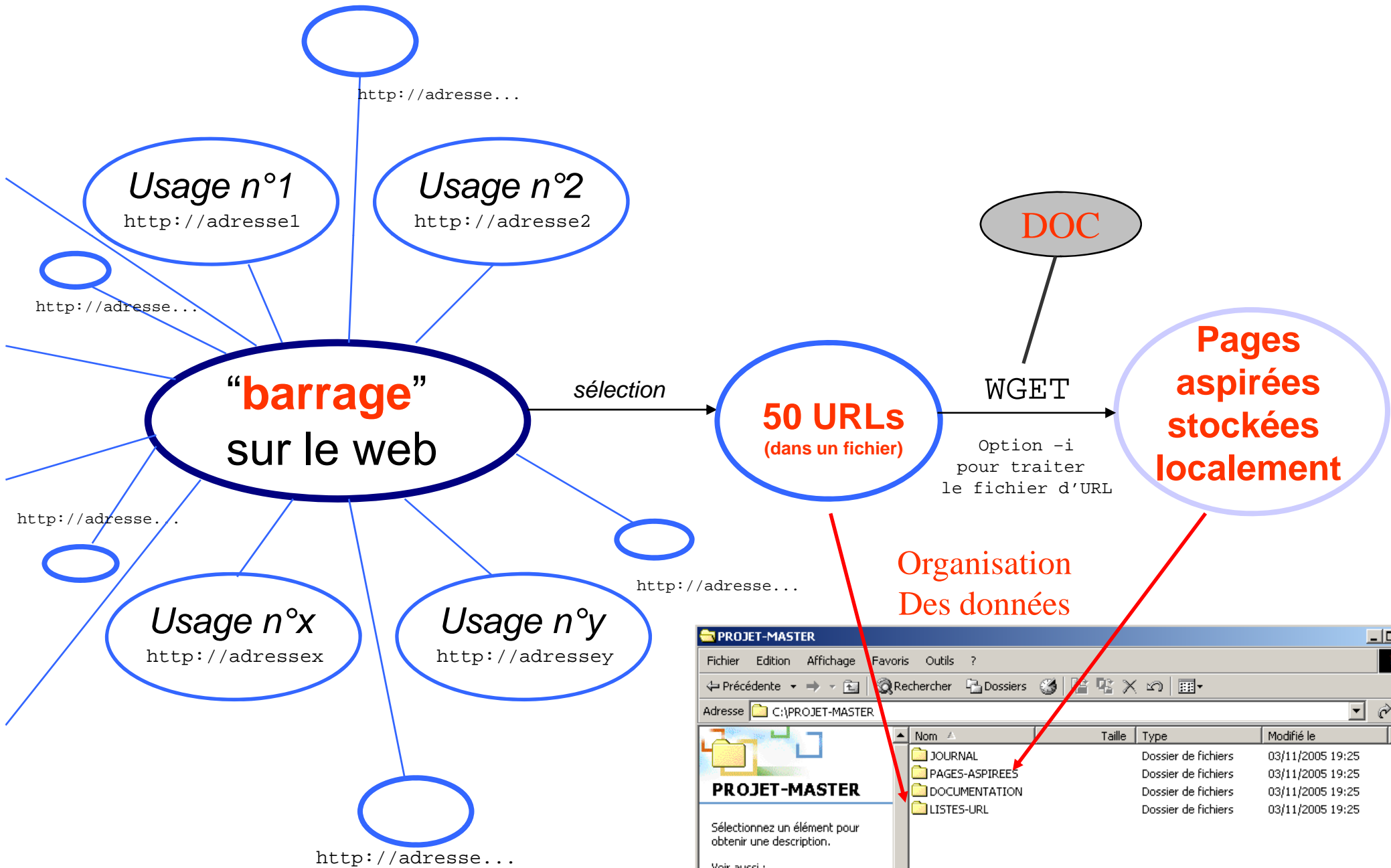


Projet n°1

- Activités : phase 2
 - Automatisation des traitements

« barrage » sur le web

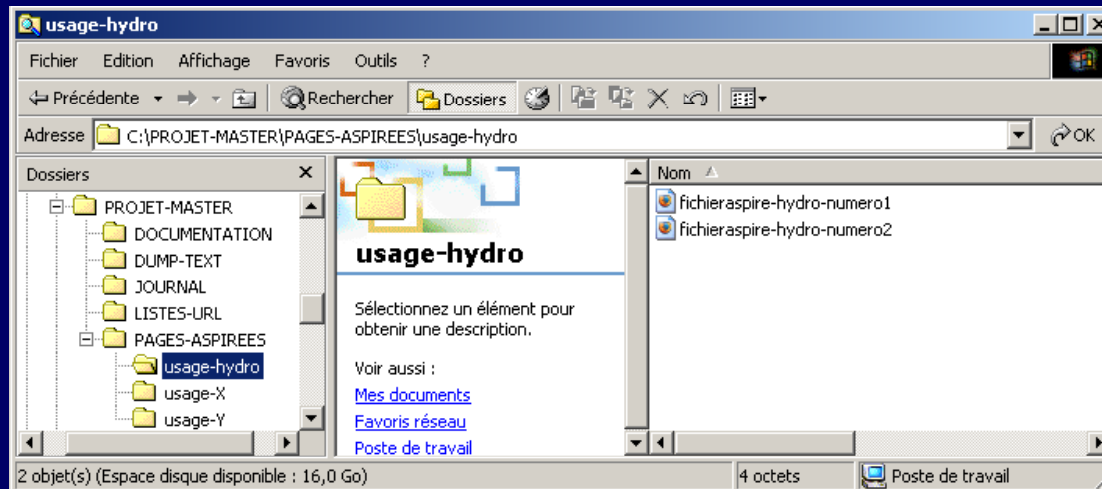
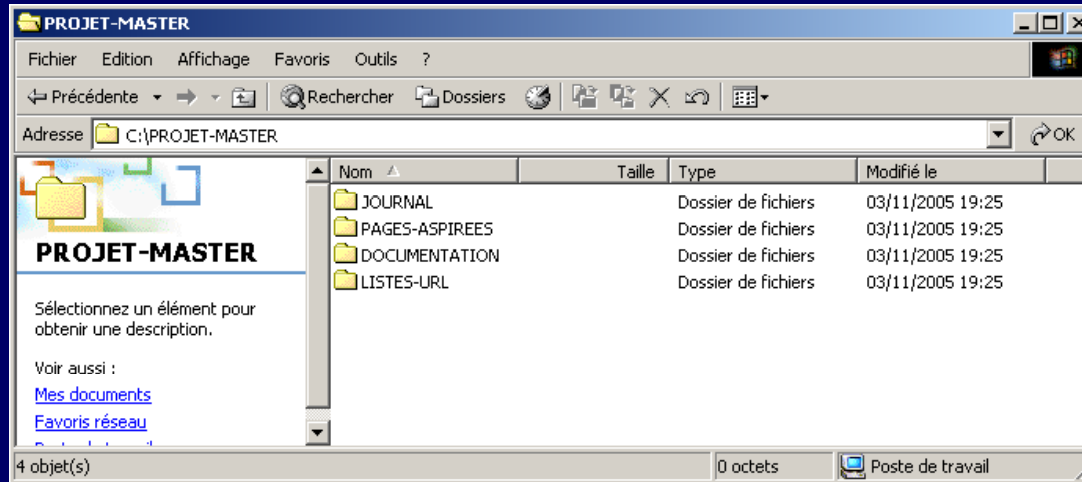
- Objectif :
 - Constituer un corpus « restreint » sur le Web (50 pages minimum) avec un échantillon des différents usages du mot « barrage »
 - Cf à partir des ressources vues la semaine dernière (sur le web, sur un corpus de Presse)
 - Outil à utiliser pour la collecte
 - wget



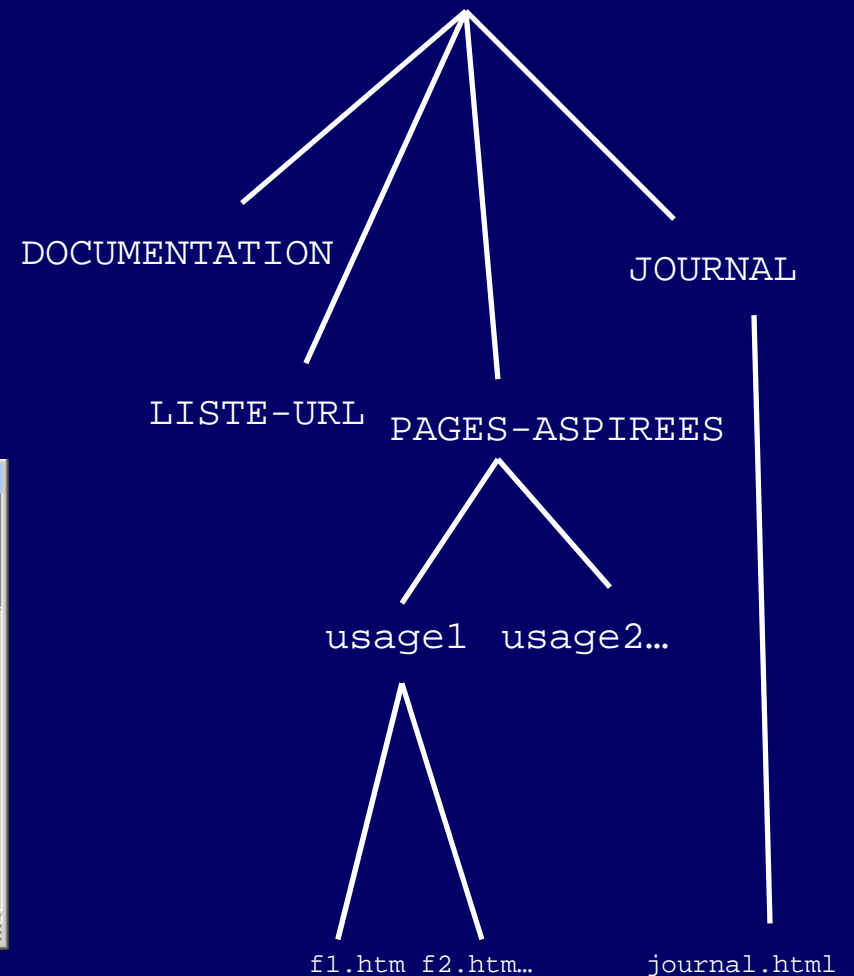
Organiser la récolte

- Organiser les données récupérées
- Sauvegarder ces données
- Rédiger des « notes » sur les données traitées
 - nom de fichier, contenu, origine, remarques etc.

Arborescence Projet (1)



C : / PROJET-MASTER



Mon journal d'activités (1)

The screenshot shows the NVU editor interface with a document titled "projet barrage". The document content includes:

- Notre projet**
- Auteurs : JMD (Inalco), SF (Paris 3), BH (Paris X)**
- Préambule**
- Tableau synthétique : les données traitées**

The table contains the following data:

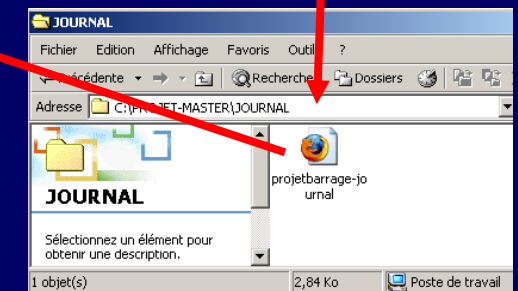
USAGE : HYDRAULIQUE	
http://www.hydro-numero1.com	fichieraspire-hydro-numero1.html
http://www.hydro-numero2.com	fichieraspire-hydro-numero2.html
etc.	etc.
USAGE : X	
http://www.usagex-numero1.com	fichieraspire-usagex-numero1.html
http://www.usagex-numero2.com	fichieraspire-usagex-numero2.html
etc.	etc.
USAGE : Y	
http://www.usagey-numero1.com	fichieraspire-usagey-numero1.html
http://www.usagey-numero2.com	fichieraspire-usagey-numero2.html
etc.	etc.

Annotations in the image:

- URL par type** (green text) points to the first column of the table.
- Fichiers aspirés** (green text) points to the second column of the table.
- Liens hypertextes vers les URLs originales** (purple text) points to the first column.
- Liens hypertextes vers les fichiers aspirés** (purple text) points to the second column.

RAPPEL :
Utilisation obligatoire de NVU pour rédiger le journal

Édition du journal avec NVU



Des chemins vers les « fichiers barrages »

- Chemin relatif : localisation par rapport à un point d'ancrage
 - Exemple :
 - ../repertoire/fichier.txt
- Chemin absolu : localisation complète d'une ressource
 - Exemple
 - C:\repertoire\fichier.txt (arborescence physique)
 - /home/etudiant/fichier.txt (arborescence logique)
 - <http://www.monsite.com>
- Dans le journal créé sous NVU :
 - Les liens vers les ressources externes (les URLs) seront des liens absolus
 - Les liens vers les ressources internes (les fichiers) seront des liens relatifs
 - NVU se charge de coder « correctement » ces 2 types de lien, la preuve...

CODAGE HTML du journal (onglet preview)

```

21. donn&eacute;es trait&eacute;es</h1>
22. <br>
23. <table style="text-align: left; width: 80%;" border="1"
24. cellpadding="2" cellspacing="2">
25.   <tbody>
26.     <tr align="center">
27.       <td style="background-color: rgb(192, 192, 192);"
28. colspan="3" rowspan="1"><span
29. style="font-weight: bold;">USAGE : HYDRAULIQUE</span></td>
30.     </tr>
31.     <tr>
32.       <td><a href="http://www.hydro-numero1.com">http://www.hydro-numero1.com</a></td>
33.       <td><a
34. href=" ../PAGES-ASPIREES/usage-hydro/fichieraspire-hydro-numero1.html">fichieraspire-hydro-numero1.html</a></td>
35.       <td><a
36. href=" ../DUMP-TEXT/fichieraspire-hydro-n1-dump.txt">fichieraspire-hydro-n1-dump.txt</a></td>
37.     </tr>
38.     <tr>
39.       <td><a href="http://www.hydro-numero2.com">http://www.hydro-numero2.com</a></td>
40.       <td><a
41. href=" ../PAGES-ASPIREES/usage-hydro/fichieraspire-hydro-numero2.html">fichieraspire-hydro-numero2.html</a></td>
42.       <td><a
43. href=" ../DUMP-TEXT/fichieraspire-hydro-n2-dump.txt">fichieraspire-hydro-n2-dump.tx</a></td>
44.     </tr>
45.     <tr>
46.       <td>etc.</td>
47.       <td>etc.</td>
48.       <td></td>
49.     </tr>
50.     <tr style="font-weight: bold;" align="center">
51.       <td style="background-color: rgb(204, 204, 204);"
52. colspan="3" rowspan="1">USAGE : X</td>
53.     </tr>
54.     <tr>
55.       <td><a href="http://www.usagex-numero1.com">http://www.usagex-numero1.com</a></td>
56.       <td><a
57. href=" ../PAGES-ASPIREES/usage-X/fichieraspire-usagex-numero1.html">fichieraspire-usagex-numero1.html</a></td>
58.       <td><a
59. href=" ../DUMP-TEXT/fichieraspire-usagex-n1-dump.txt">fichieraspire-usagex-n1-dump.txt</a></td>
60.     </tr>
61.     <tr>
62.       <td><a href="http://www.usagex-numero2.com">http://www.usagex-numero2.com</a></td>
63.       <td><a
64. href=" ../PAGES-ASPIREES/usage-X/fichieraspire-usagex-numero2.html">fichieraspire-usagex-numero2.html</a></td>
65.       <td><a
66. href=" ../DUMP-TEXT/fichieraspire-usagex-n2-dump.txt">fichieraspire-usagex-n2-dump.txt</a></td>
67.     </tr>
68.     <tr>
69.       <td>etc.</td>
70.       <td>etc.</td>
71.       <td></td>
72.     </tr>
73.     <tr style="font-weight: bold;" align="center">
74.       <td style="background-color: rgb(204, 204, 204);"
75. colspan="3" rowspan="1">USAGE : Y</td>

```

chemin absolu

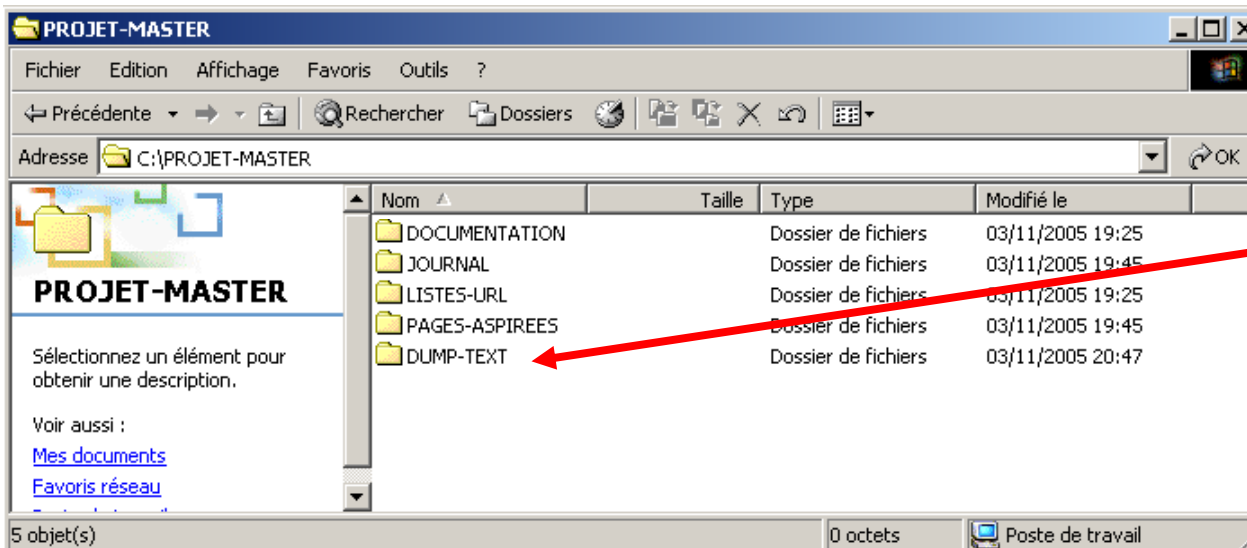
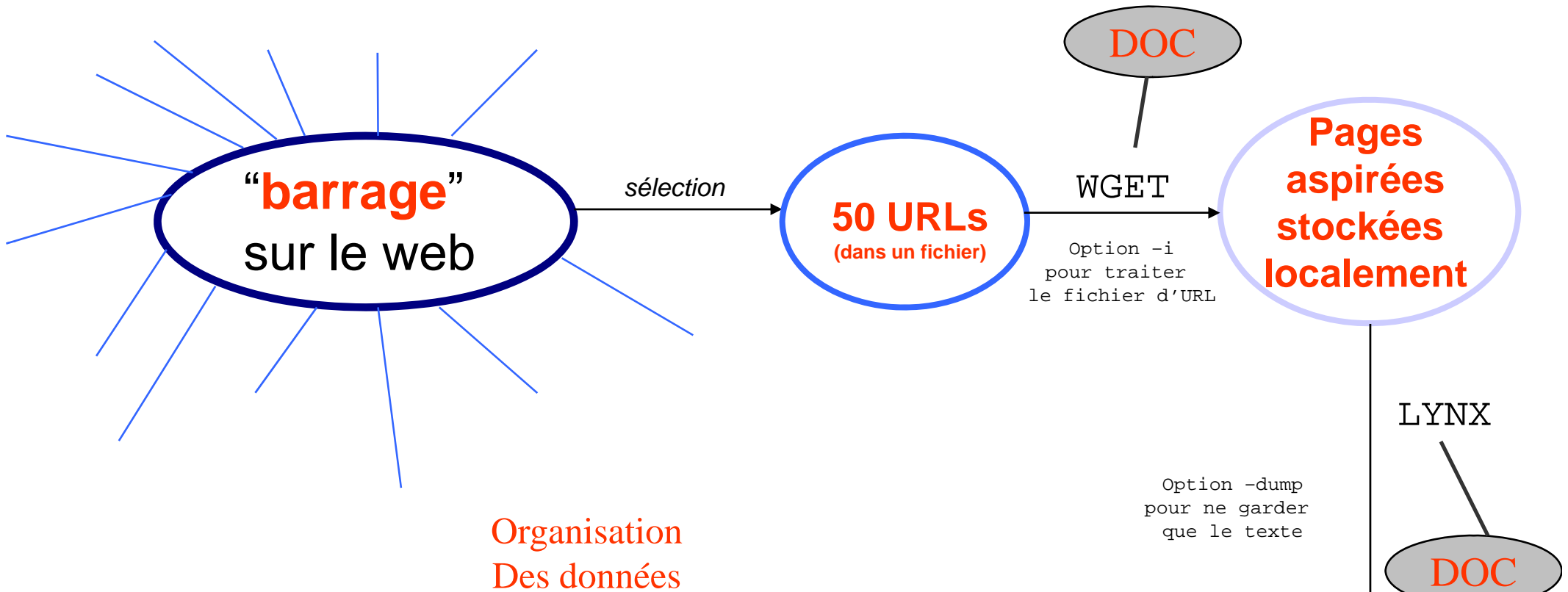
chemin relatif

WGET : mode d'emploi (1)

- Rappel : syntaxe commande Unix
 - commande [options] paramètres
 - En général, les options sont précédées du symbole - et peuvent être groupées
 - `ls -l /home/etudiant/`
 - `ps -ae`
 - Les paramètres précisent les fichiers concernés
 - `wget` est une commande unix
 - Syntaxe similaire : `cf man`

WGET : mode d'emploi (2)

- « barrage » aux options de wget
 - wget <http://fly.srk.fer.hr/>
 - Utilisation simple
 - wget -i <file>
 - le fichier <file> contient les URL que vous voulez télécharger
 - wget -r <http://fly.srk.fer.hr/>
 - Récursion



Mon journal d'activités (2)

RAPPEL :

Utilisation obligatoire de NVU pour rédiger le journal

Édition du journal avec NVU

Notre projet

Auteurs : JMD (Inalco), SF (Paris 3), BH (Paris X)

Préambule
 blah
 blah

Tableau synthétique : les données traitées

USAGE : HYDRAULIQUE		
http://www.hydro-numero1.com	fichieraspire-hydro-numero1.html	fichieraspire-hydro-n1-dump.txt
http://www.hydro-numero2.com	fichieraspire-hydro-numero2.html	fichieraspire-hydro-n2-dump.txt
etc.	etc.	
USAGE : X		
http://www.usagex-numero1.com	fichieraspire-usagex-numero1.html	fichieraspire-usagex-n1-dump.txt
http://www.usagex-numero2.com	fichieraspire-usagex-numero2.html	fichieraspire-usagex-n2-dump.txt
etc.	etc.	
USAGE : Y		
http://www.usagey-numero1.com	fichieraspire-usagey-numero1.html	fichieraspire-usagey-n1-dump.txt
http://www.usagey-numero2.com	fichieraspire-usagey-numero2.html	fichieraspire-usagey-n2-dump.txt
etc.	etc.	

URL par type

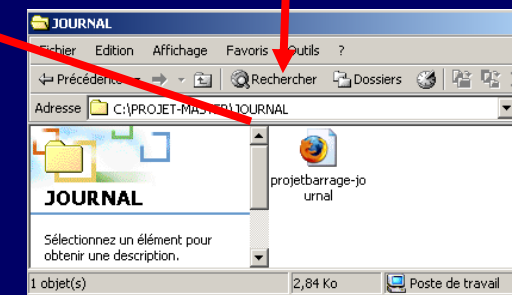
Fichiers aspirés

Fichiers « dump »

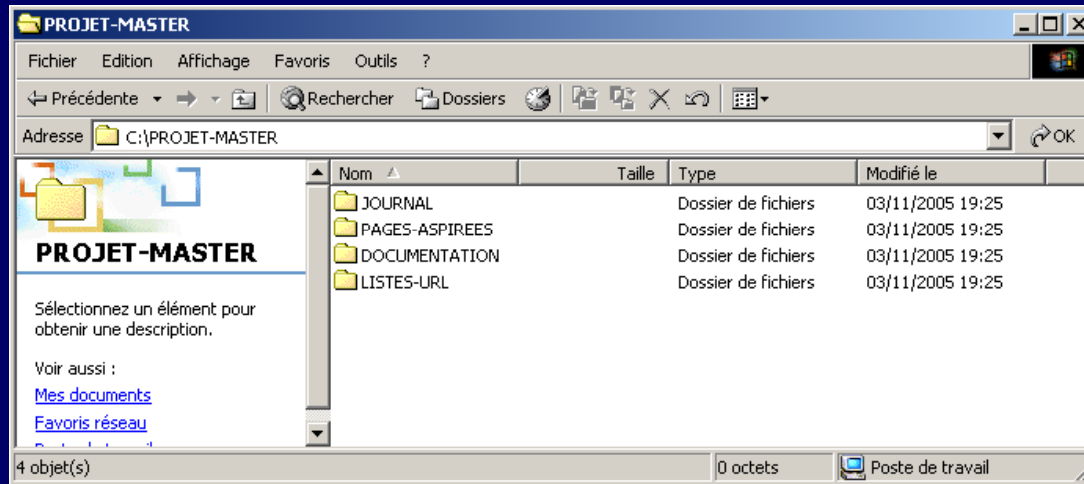
Liens hypertextes vers les URLs originales

Liens hypertextes vers les fichiers aspirés

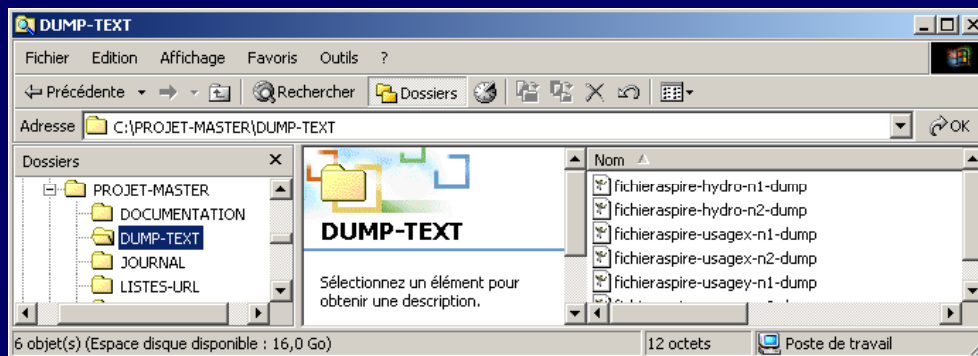
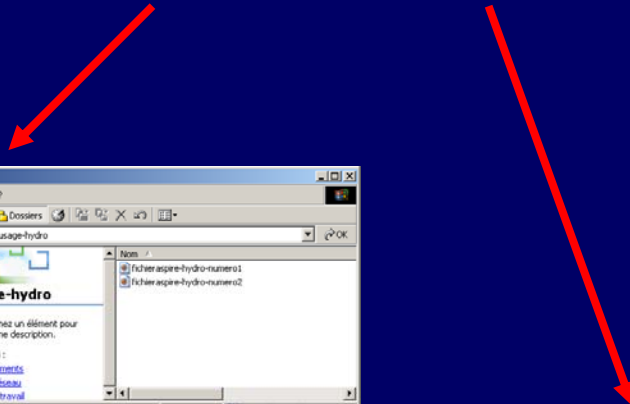
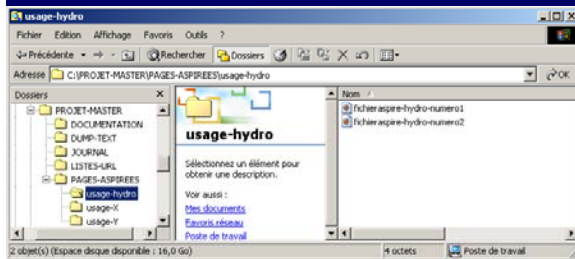
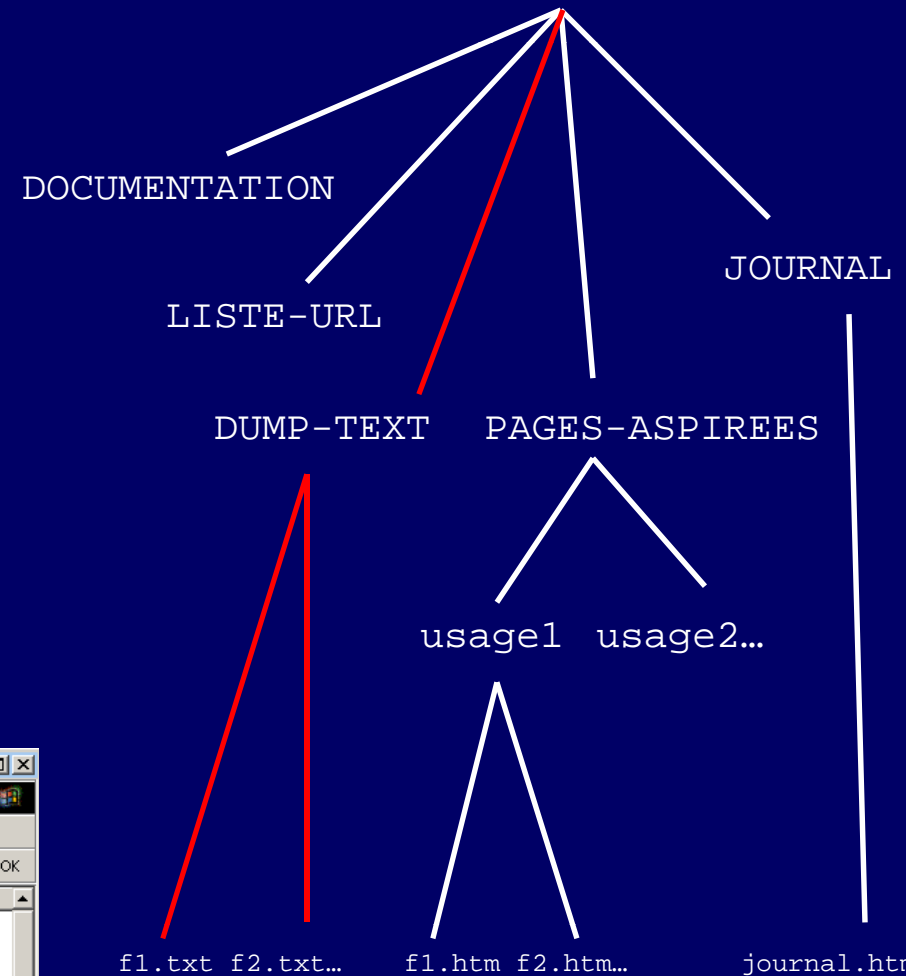
Liens hypertextes vers les fichiers « dumpés »



Arborescence Projet (2)



C : / PROJET-MASTER



f1.txt f2.txt... f1.htm f2.htm... journal.html

LYNX : navigateur en ligne de commandes

- Lynx est un petit client web/news/ftp. Il existe sous UNIX, sous DOS, et même sous Windows...
 - Lynx un cousin de Internet Explorer ou Firefox
- Qu'a-t-il de plus que votre « Mircoscape »
« Explocommunicateur » habituel ?
 - Il fonctionne en **mode texte**. Ça peut paraître ridicule voire désuet, c'est parfois très utile...

LYNX : mode d'emploi (1)

- Lancer la fenêtre de commande, puis taper « lynx » : rustique !!!!!



```
Sélectionner file://localhost/c:/
lynx32 directory <pl of 2>
lynx32
Up to c:
Subdirectories:
  Oct 08 14:06 help
  Oct 08 14:06 icon
  Oct 08 14:06 samples
Files:
  18K Mar 14 1997 COPYING
  1K Mar 02 1997 HELPER.TXT
  1K Sep 05 2004 lynx.bat
  1K Dec 01 2002 lynx.bat~
  129K Sep 05 2004 lynx.cfg
  129K Dec 01 2002 lynx.cfg~
  1135K Jul 19 2000 lynx.exe
  30K Jul 19 2000 lynx.man
-- press space for next page --
Arrow keys: Up and Down to move. Right to follow a link; Left to go back.
H>help O>ptions P>rint G>o M>ain screen Q>uit /=search [delete]=history list
```

LYNX : mode d'emploi (2)

- Naviguons avec Lynx : `lynx URL`
 - Ex : `lynx http://tal.univ-paris3.fr/plurital/`

```

http://tal.univ-paris3.fr/plurital/index.html
pluriTAL : Filieres TAL et ingénierie linguistique de Paris I.. <p1 of 4>

PluriTAL

AccueilActualitésAdministrationMaster
PLURITALJournées-CoursLiensContactBibliographieLecturesGroupe
PluriTALImages

Filieres TAL et ingénierie linguistique de Paris III Sorbonne nouvelle,
Paris X Nanterre, INALCO (Institut National des langues et civilisations
orientales)

Lieux/Sites PluriTAL

Localisations de Plurital (comment y aller)
Site Web officiel : http://tal.univ-paris3.fr/plurital/
Site Web miroir : http://www.cavi.univ-paris3.fr/ilpga/plurital/

Rentrée 2005-2006

Journée d'accueil du MASTER le 10 octobre 2005, 14h00-16h00, dans les
salons de l'INALCO.
-- press space for next page --
Arrow keys: Up and Down to move. Right to follow a link; Left to go back.
H)elp O)ptions P)rint G)o M)ain screen Q)uit /=search [delete]=history list

```

LYNX : mode d'emploi (3)

- « Ce qui est bon dans le lynx »
 - Lynx permet de filtrer / « dumper » le texte !!!
 - `lynx -dump http://tal.univ-paris3.fr/plurital/`

```

Cygwin B20
bash-2.02$ lynx -dump http://tal.univ-paris3.fr/plurital/

PluriTAL

[1]Accueil[2]Actualités[3]Administration[4]Master
PLURITAL[5]Journées-Cours [6]Liens [7]Contact [8]Bibliographie [9]Lectures
[10]Groupe PluriTAL[11]Images

Filières TAL et ingénierie linguistique de Paris III Sorbonne nouvelle,
Paris X Nanterre, INALCO (Institut National des langues et civilisations
orientales)

Lieux/Sites PluriTAL

[12]Localisations de Plurital (comment y aller)
Site Web officiel : [13]http://tal.univ-paris3.fr/plurital/
Site Web miroir : [14]http://www.cavi.univ-paris3.fr/ilpga/plurital/

Rentrée 2005-2006

Journée d'accueil du [15]MASTER le 10 octobre 2005, 14h00-16h00, dans
les salons de l'INALCO.

PluriTAL - Apports croisés, complémentaires et pluriels pour le TAL

Le domaine du TAL et de l'ingénierie linguistique se caractérise par
la multiplicité des dimensions prises en compte (traitement de
l'écrit, de l'oral), des niveaux d'analyse impliqués (morphologie,
syntaxe, sémantique, pragmatique), des techniques, des langues
abordées. C'est un domaine où il est essentiel de conjuguer des


```

« Problème » :
Affichage sur la
sortie standard
i.e l'écran

Nécessité de rediriger
le flux de sortie...

LYNX : mode d'emploi (4)

- lynx -dump <http://tal.univ-paris3.fr/plurital/> > resultat.txt



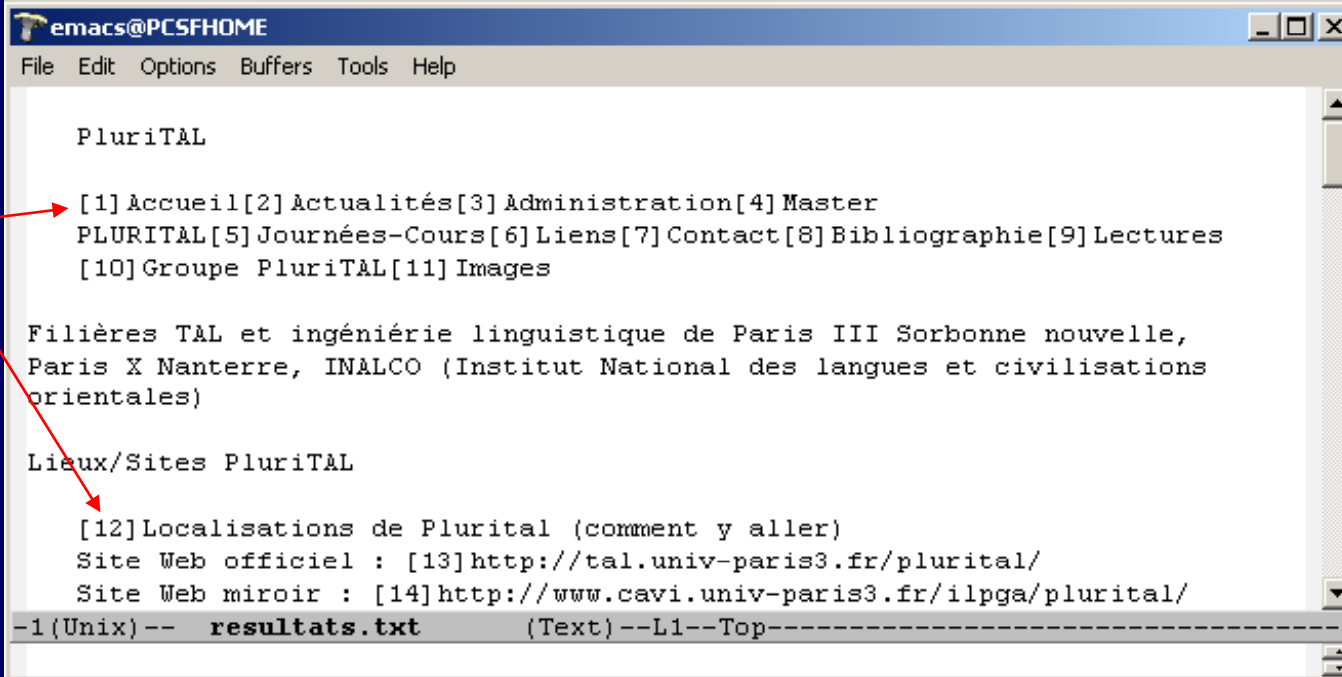
On indique dans la ligne de commande que le flux de sortie de la commande est « envoyé » dans un fichier

```

Cygwin B20
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$
bash-2.02$ lynx -dump http://tal.univ-paris3.fr/plurital/ > resultat.txt
bash-2.02$
  
```

L'URL « dumpée » est sauvegardée au format texte dans le fichier resultat.txt

Les liens sont numérotés
Et listés en fin de fichier
Nettoyage à prévoir...



```

emacs@PCSFHOME
File Edit Options Buffers Tools Help

PluriTAL

[1] Accueil[2] Actualités[3] Administration[4] Master
PLURITAL[5] Journées-Cours[6] Liens[7] Contact[8] Bibliographie[9] Lectures
[10] Groupe PluriTAL[11] Images

Filières TAL et ingénierie linguistique de Paris III Sorbonne nouvelle,
Paris X Nanterre, INALCO (Institut National des langues et civilisations
orientales)

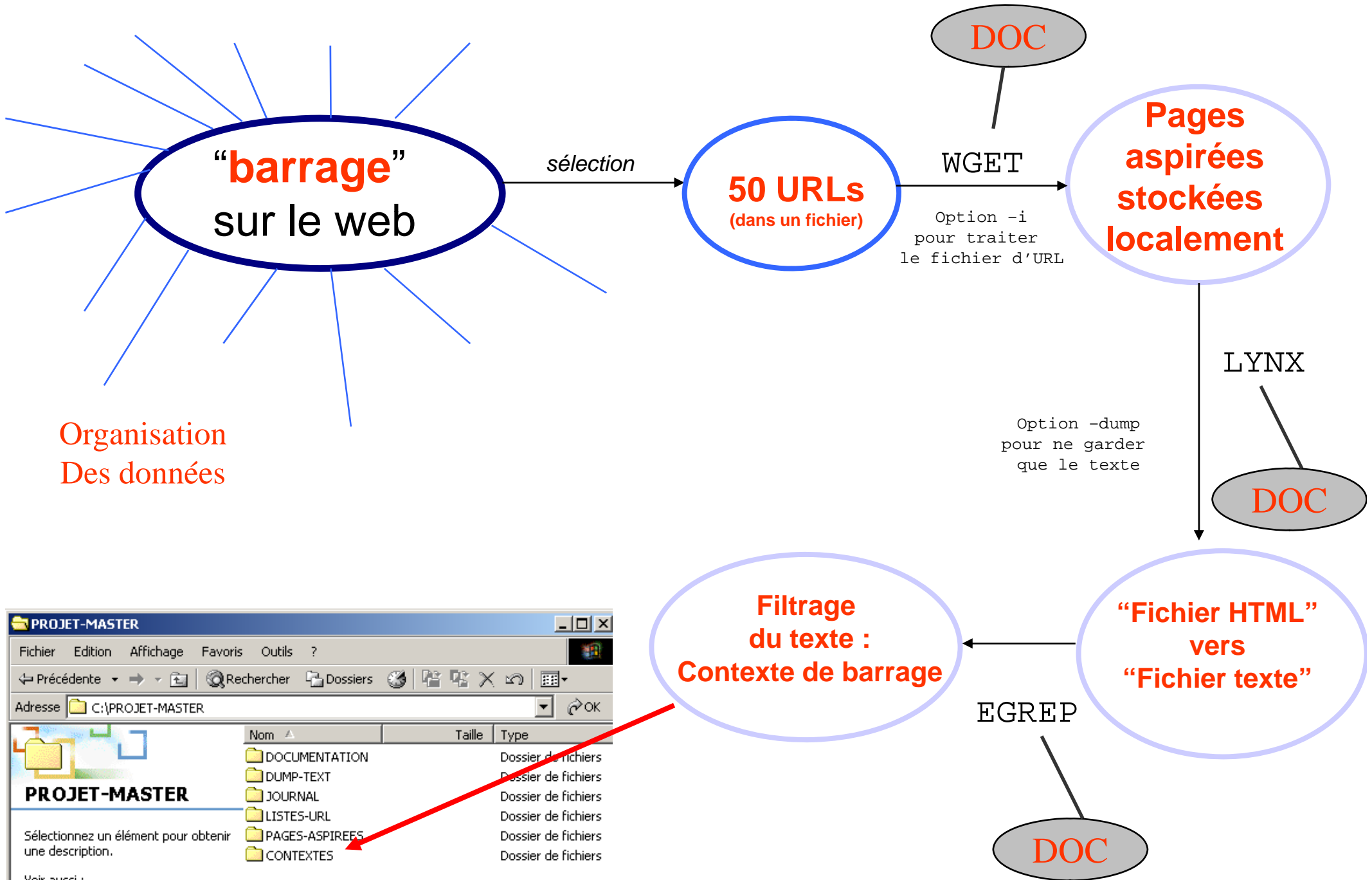
Lieux/Sites PluriTAL

[12] Localisations de Plurital (comment y aller)
Site Web officiel : [13]http://tal.univ-paris3.fr/plurital/
Site Web miroir : [14]http://www.cavi.univ-paris3.fr/ilpga/plurital/

-1(Unix) -- resultats.txt (Text) --L1--Top--
  
```

LYNX : utilisation

- Lynx est disponible avec la version de cygwin installée...



Mon journal d'activités (3)

URL
par
type

Fichiers
aspirés

Fichiers
« dump »

Contextes

Notre projet

Auteurs : JMD (Inalco), SF (Paris 3), BH (Paris X) *Liens hypertextes vers les URLs originales*

Préambule
 blah
 blah

Liens hypertextes vers les fichiers aspirés

Liens hypertextes vers les fichiers « dumpés »

Liens hypertextes vers les contextes

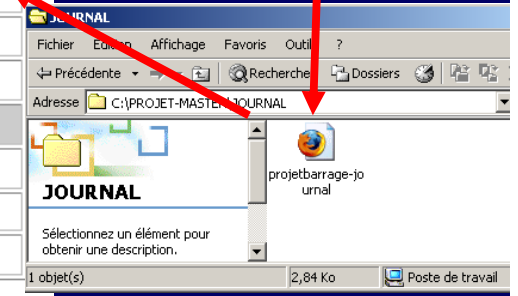
Tableau synthétique : les données traitées

USAGE : HYDRAULIQUE			
http://www.hydro-numero1.com	fichieraspire-hydro-numero1.html	fichieraspire-hydro-n1-dump.txt	contexte-hydro-1.txt
http://www.hydro-numero2.com	fichieraspire-hydro-numero2.html	fichieraspire-hydro-n2-dump.txt	contexte-hydro-2.txt
etc.	etc.		
USAGE : X			
http://www.usagex-numero1.com	fichieraspire-usagex-numero1.html	fichieraspire-usagex-n1-dump.txt	contexte-usagex-1.txt
http://www.usagex-numero2.com	fichieraspire-usagex-numero2.html	fichieraspire-usagex-n2-dump.txt	contexte-usagex-2.txt
etc.	etc.		
USAGE : Y			
http://www.usagey-numero1.com	fichieraspire-usagey-numero1.html	fichieraspire-usagey-n1-dump.txt	contexte-usagey-1.txt
http://www.usagey-numero2.com	fichieraspire-usagey-numero2.html	fichieraspire-usagey-n2-dump.txt	contexte-usagey-2.txt
etc.	etc.		

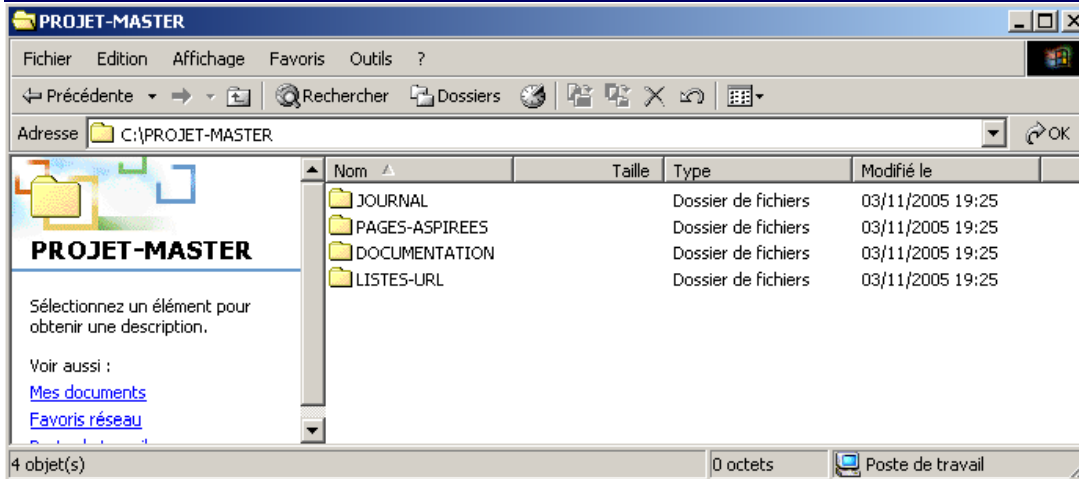
RAPPEL :

Utilisation obligatoire de NVU pour rédiger le journal

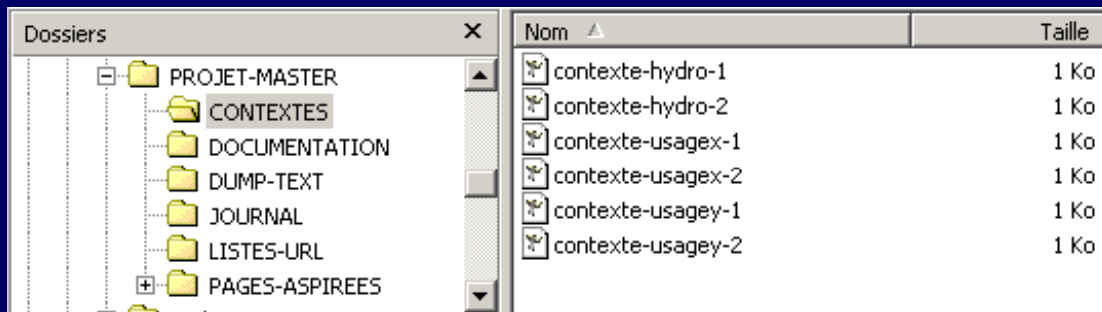
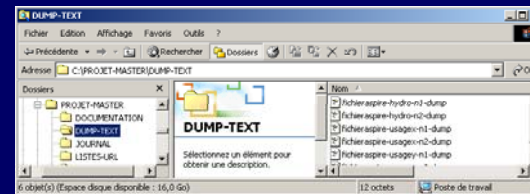
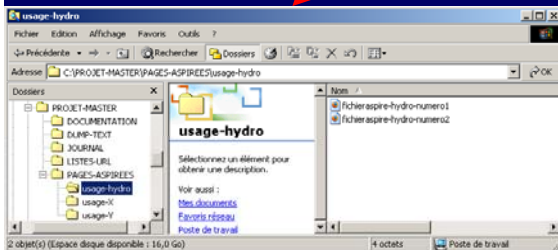
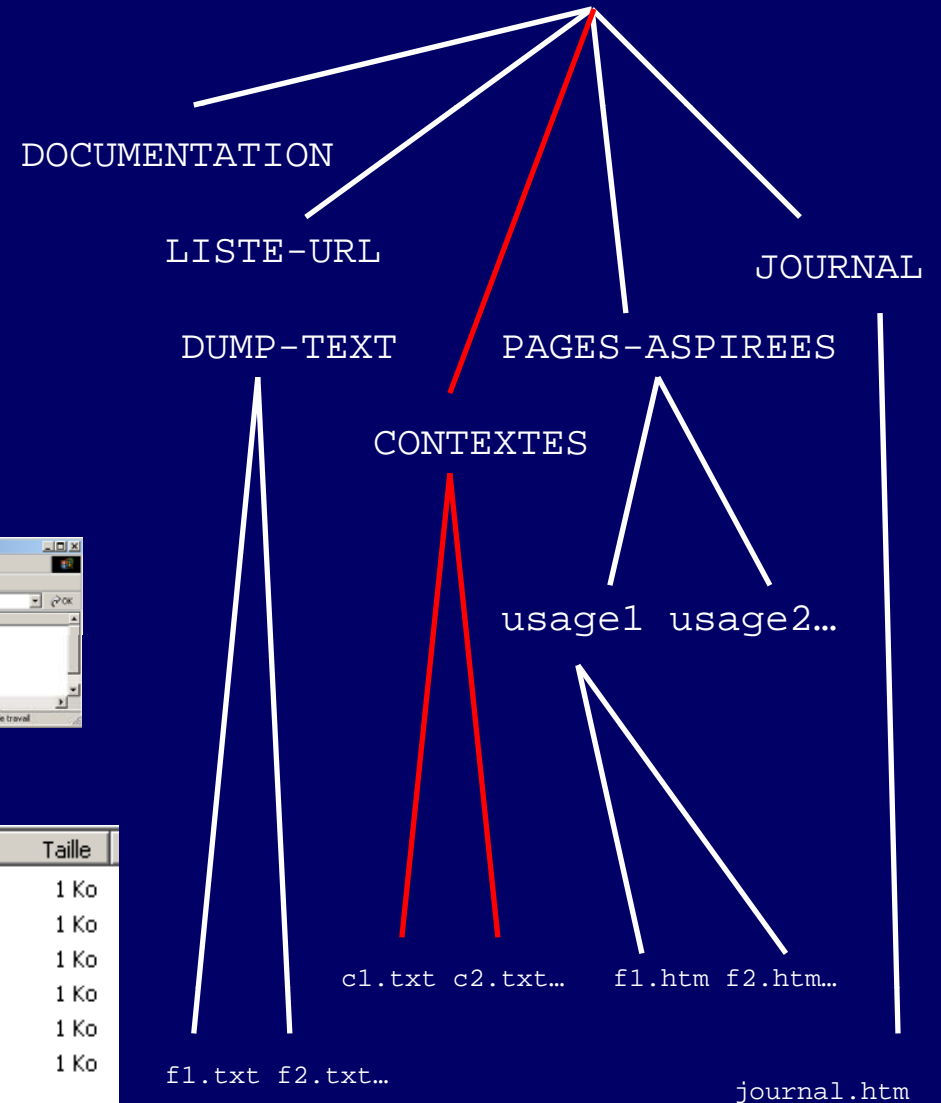
Édition du journal avec NVU



Arborescence Projet (3)



C : / PROJET-MASTER



Filterer la récolte

- A l'issue de la récolte, vous disposez
 - De données au format HTML...
 - De données au format TEXTE (dump)
- On ne souhaite conserver que des contextes larges du mot « barrage » dans ces fichiers...
- Filtrage de la récolte
 - `egrep` : commande unix
 - Les expressions régulières (nécessaires pour `egrep`)

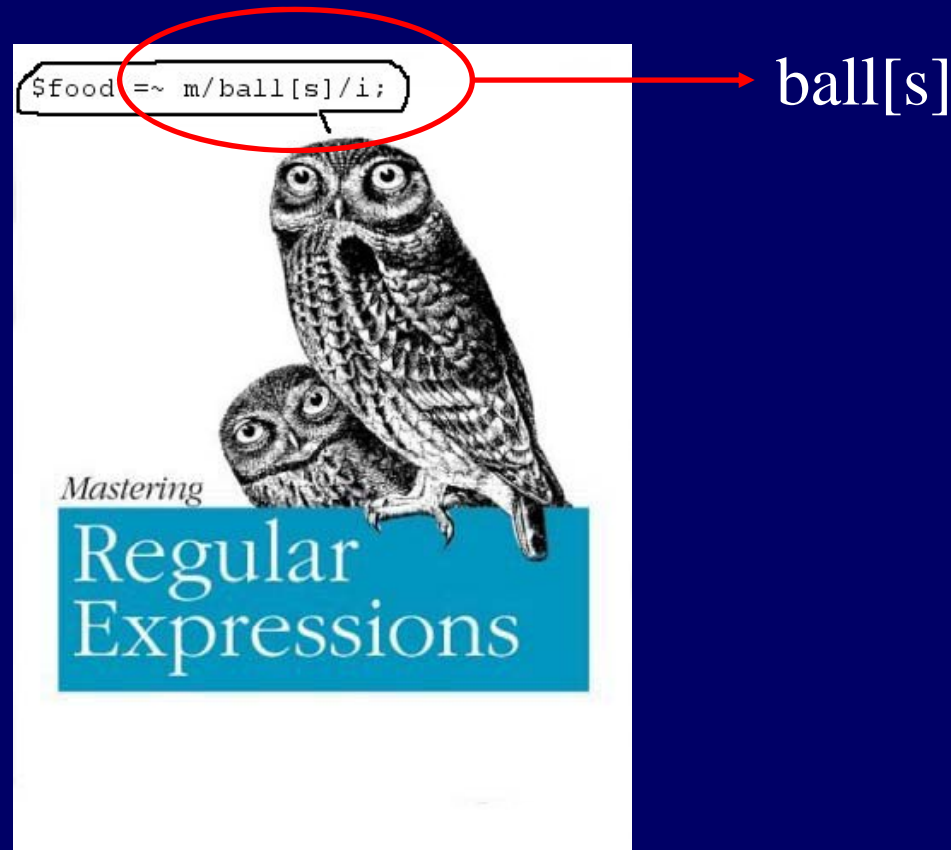
egrep

- Cet utilitaire permet de rechercher dans des fichiers des lignes contenant un motif donné. Son utilisation implique que l'on maîtrise les expressions régulières.
- Le format d'une commande est la suivante :
 - `egrep <motif> <fichier>`

Options egrep

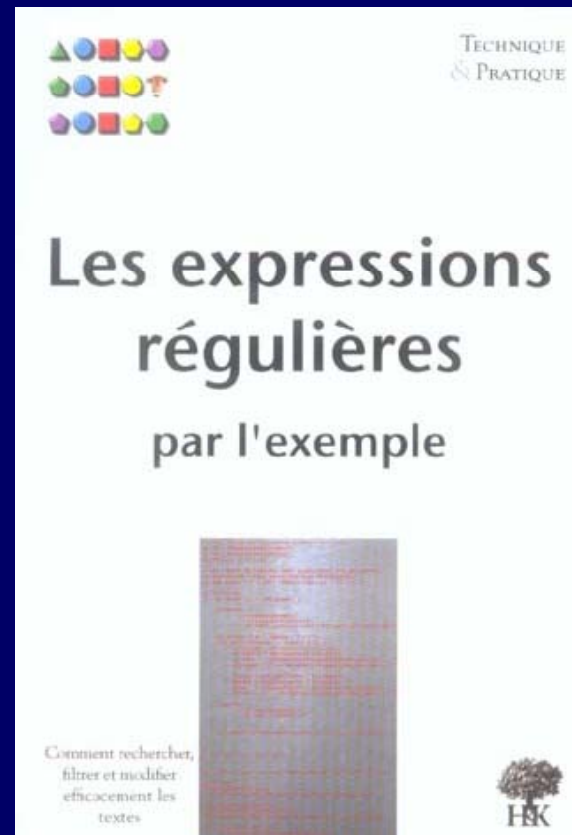
- `egrep` peut être utilisé avec des options qui permettent de modifier son comportement :
 - `-c` : affiche un décompte des lignes comprenant le motif cherché.
 - `-v` : affiche les lignes qui ne contiennent pas le motif.
 - `-n` : chaque ligne qui contient le motif est précédée de son numéro dans le fichier.
 - `-i` : permet de ne pas tenir compte de la différence entre minuscules et majuscules
 - `-A nbx`, `-B nby` : pour récupérer `nbx` lignes de contexte après (AFTER), et `nby` lignes de contexte avant (BEFORE)
 - Etc. (*cf man*)

- Les expressions régulières
 - Les expressions régulières permettent de représenter de manière générique des motifs textuels



Biblio (courte)

- *Les Expressions Régulières par L'exemple, Fourmond, Vincent, Editeur : H & K, 2005*



Les méta-caractères

- Appelés aussi caractères spéciaux, ce sont des caractères interprétés *en contexte expression rationnelle* comme des opérateurs.
- En voici la liste avec un bref descriptif :
 - . (point) représente un caractère qcq, sauf \n
 - * (astérisque) répétition du caractère précédent
 - + au moins une occurrence de l'expression régulière
 - ? au plus une occurrence de l'expression régulière
 - [...] (crochets) l'un des caractères de l'ensemble.
 - [^..] en début de crochets recherche dans le complémentaire de l'ensemble
 - ^ recherche en début de ligne
 - \$ recherche en fin de ligne
 - \ annule le rôle de méta-caractère, pour jouer le rôle du caractère usuel
 - {n,m} indique le nombre de répétitions attendus du caractère précédent
 - | joue le rôle de "ou" entre 2 expr rég.
- L'antislash \ inhibe l'interprétation des caractères spéciaux et force leur interprétation usuelle.
 - Exemples
 - `.\txt` recherche les chaînes du genre `c.txt`, où `c` est un caractère unique qcq
 - `*$` recherche les chaînes qui se terminent (\$) par le caractère astérisque (*)

Expressions régulières simples

- Soit *expat* une *expression régulière atomique* quelconque, alors ce sont
 - **expat1expat2expat3 ...** toute concaténation sans espace formée d'expr. régulière atomiques
 - **expat*** chaîne composée de 0 à N caractères vérifiant *expat*
 - **expat+** chaîne contenant au moins un caractère vérifiant *expat*
 - **expat?** chaîne contenant au plus un caractère vérifiant *expat*
 - **expat{n}** chaîne composée exactement de n caractères vérifiant *expat*
 - **expat{n,}** chaîne composée d'au moins n caractères vérifiant *expat*
 - **expat{n,m}** chaîne composée de n à m caractères vérifiant *expat*

■ Exemples

- a^* caractère de multiplication, suivant un caractère signifie la répétition de 0 à plusieurs exemplaires de ce caractère.
- $[ab]^*$ signifie répétition possible d'un quelconque des 2 caractères a ou b.
- a^+ répétition de 1 à plusieurs exemplaires de a
- $a?$ 0 ou 1 exemplaire de a
- $c[ad]^*r$ impose la présence de cr ou car ou cdr, et rien de plus.
- $/x\{5,10\}$ 5 à 10 répétitions attendues de x
- $a\{5,\}$ 5 ou plus répétitions attendues de a
- $a\{5\}$ exactement 5 occurrences de a
- $a.\{5\}b$ recherche les mots contenant 5 caractères entre a et b
- $\backslash s^+$ recherche un ou plusieurs caractères séparateurs.
- Conséquence : $\{0,\}$ $\{1,\}$ $\{0,1\}$ correspondent à $*$, $+$, $?$

Expressions régulières étendues

- Soit *expsim* une *expression régulière simple* quelconque, comme définie précédemment, alors ce sont des expressions régulières générales :
 - **expsim1expsim2 ...** toute concaténation formée d'expr. régulière simples
 - **^expsim** recherche le motif *expsim* en **début** de chaîne
 - **expsim\$** recherche le motif *expsim* en **fin** de chaîne
 - **expsim\b** le motif *expsim* doit se trouver la **fin d'un mot**
 - **expsim1|expsim2|...** alternative : recherche de *expsim1* ou *expsim2* ...
 - **(expsim)*** chaîne composée de 0 à N caractères vérifiant *expsim*
 - **(expsim)+** chaîne contenant au moins un caractère vérifiant *expsim*
 - **(expsim)?** chaîne contenant au plus un caractère vérifiant *expsim*
 - **(expsim){n}** chaîne contenant la concaténation d'exactly n chaînes vérifiant *expsim*
 - **(expsim){n,m}** chaîne composée de n à m caractères vérifiant *expsim*

■ Remarques

- Les symboles `^`, `$` sont appelés *motifs d'ancrage*, car ils permettent de préciser l'endroit où doit être effectuée la recherche du motif dans la chaîne (alors que sans directive particulière la recherche de correspondance du motif avec la chaîne s'effectue de gauche à droite de la chaîne).
- Le symbole d'ancrage `\b` indique une limite de mot obligatoire, à l'endroit où il est inséré dans le motif
- Par exemple :
 - `info\b` est vérifié par "l'info pour tous", "info-matin" mais pas par infos, informatique ...
 - `\binfo` est vérifié par "informatique", "m'informer" mais pas désinformation, ..
- Les parenthèses autour de `expsim` sont indispensables, sinon les opérateurs de répétition s'appliquent au dernier caractère de l'expression (il s'agit d'une question de priorité des opérateurs ..)
 - Par exemple, les motifs `li(nux)*` et `linux*` reconnaissent respectivement li, linux, linuxnux,.. et linu, linux, linuxx, linuxxx,...
- Attention, ces extensions ne sont pas toutes reconnues par l'ensemble des filtres. Par exemple la commande `grep`, ne connaît pas `|`, ni les parenthèses

■ Exemples

- `color|couleur` réussit si l'un des motifs, soit `color`, soit `couleur`, est trouvé
- `5[0-9]{2}` reconnaît tout nombre de 500 à 599.
- `/^(eleve|prof|stage)[3-7]/` est satisfait par les chaînes commençant par l'un des 3 mots suivis immédiatement par un numéro de 3 à 7
- comment reconnaître un mot de 6 lettres formé des lettres (cela ne vous rappelle rien ?)

■ Parenthèses de mémorisation

- La présence de parenthèses permet de mémoriser une ou plusieurs parties de l'expression qu'elle entoure, sans modifier son interprétation. Pour rappeler ces valeurs mémorisées, on utilise les notations `\1`, `\2` ... qui correspondent aux valeurs reconnues de même ordre.
- Ainsi `eleve(\d).+station\1` sera satisfait par *eleve3 au poste3* et non par *eleve5 au poste3*

() parenthésage

- Les parenthèses permettent d'indiquer comment les éléments de l'expression doivent être groupés.
- Les opérateurs s'organisent selon des niveaux de priorité :
 - fermeture > concaténation > union

Exemple : $a|bc^*d$

- On commence par s'intéresser aux signes *
- On les regroupe avec leur opérande :
 - $a|b(c^*)d$
- Ensuite on s'intéresse aux concaténations.
 - 1^{ère} concaténation : $a|(b(c^*))d$
 - 2^{ème} concaténation : $a|((b(c^*))d)$
 - Enfin, on regarde les unions :
 - $(a|((b(c^*))d))$

Fin des moissons

- Remise des données par mail
- Prochaine étape : les BOITES à OUTILS...