Programmation et projet encadré - L7TI005

Attendus du projet (version abrégée)

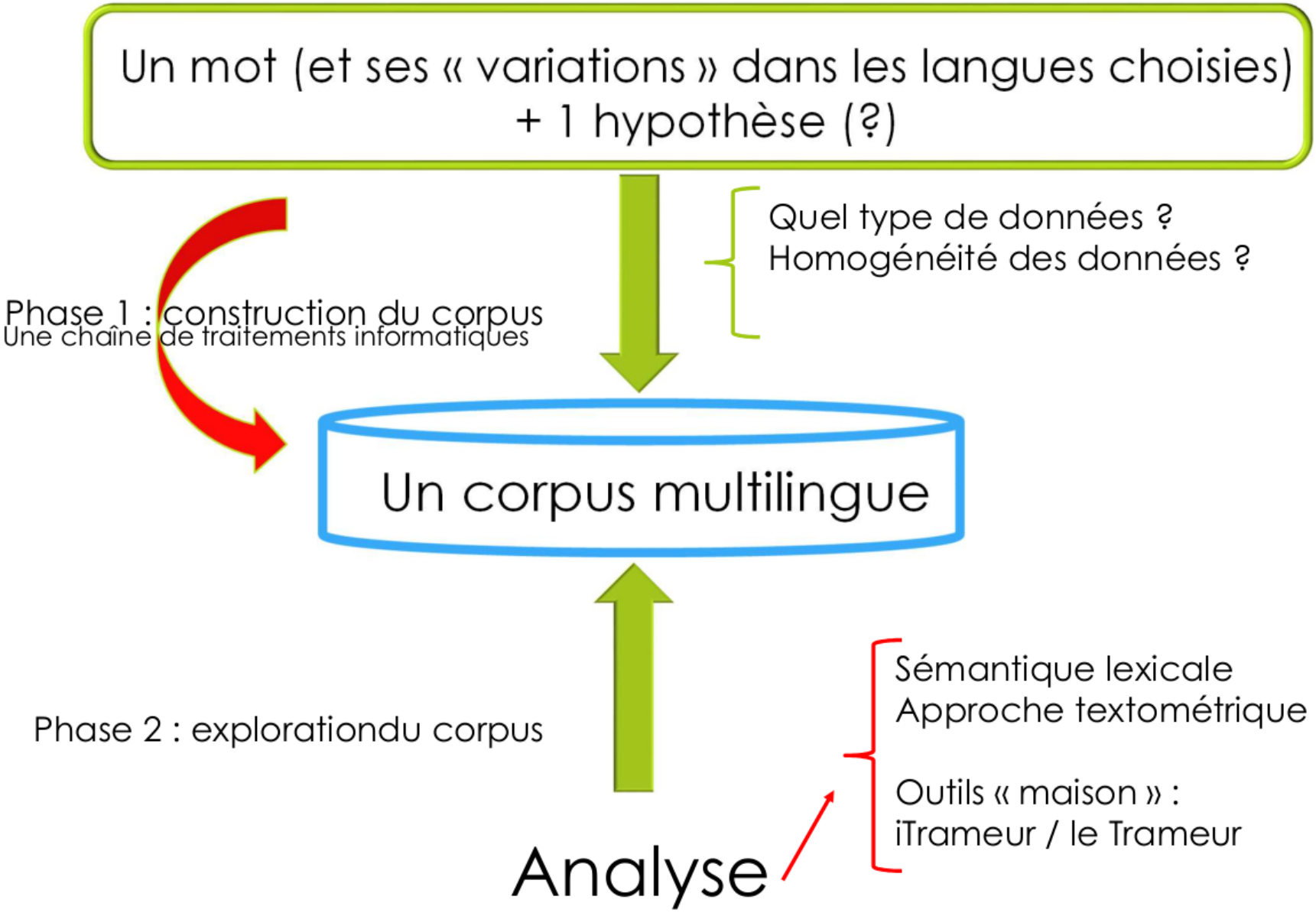
Crédits supports : Serge Fleury

Yoann Dupont, Serge Fleury prenom.nom@sorbonne-nouvelle.fr
Pierre Magistry pierre.magistry@inalco.fr

2022-2023

Université Sorbonne-Nouvelle
INALCO
Université Paris-Nanterre

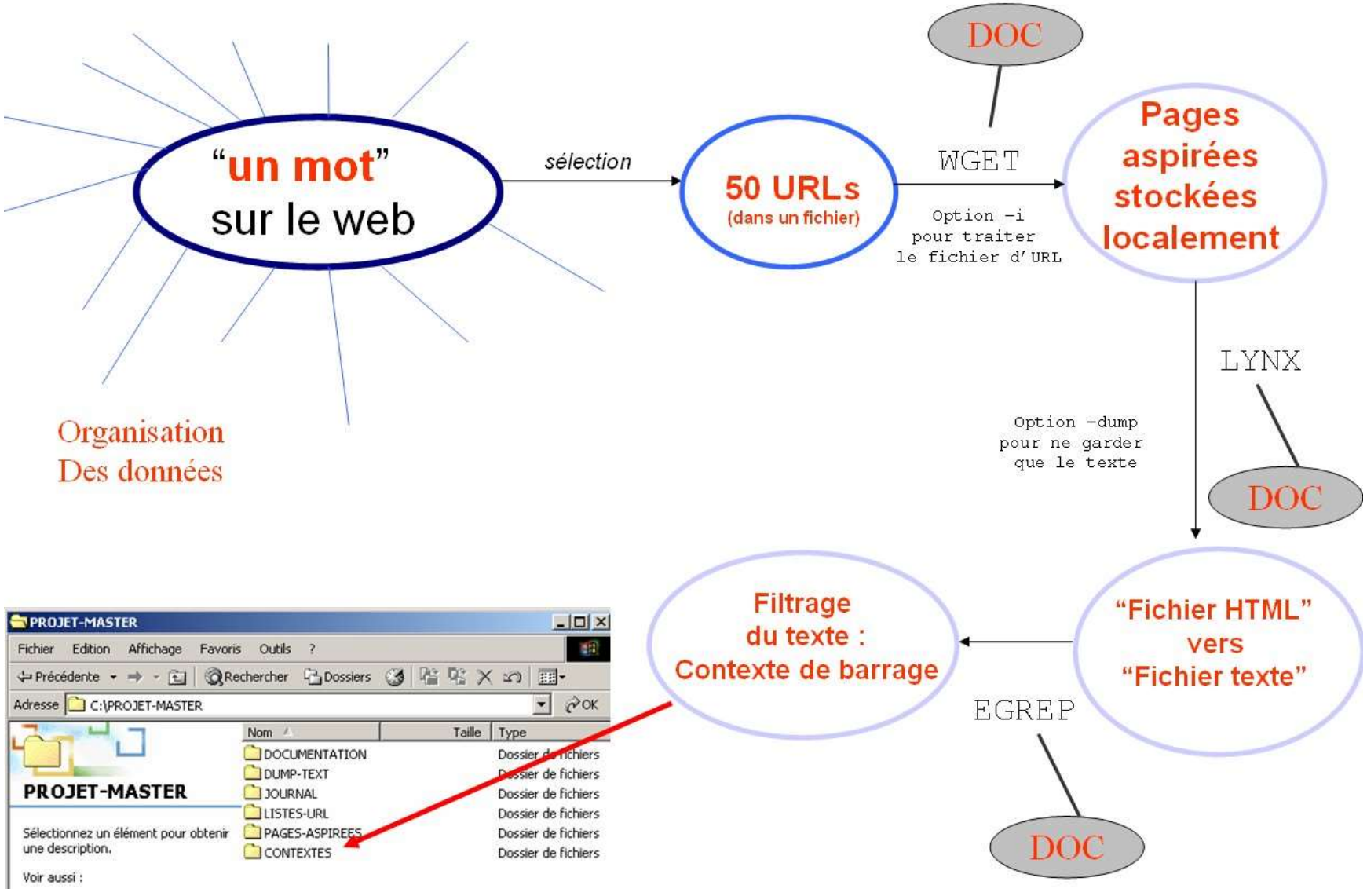
Le projet : Description de la démarche à mettre en œuvre...

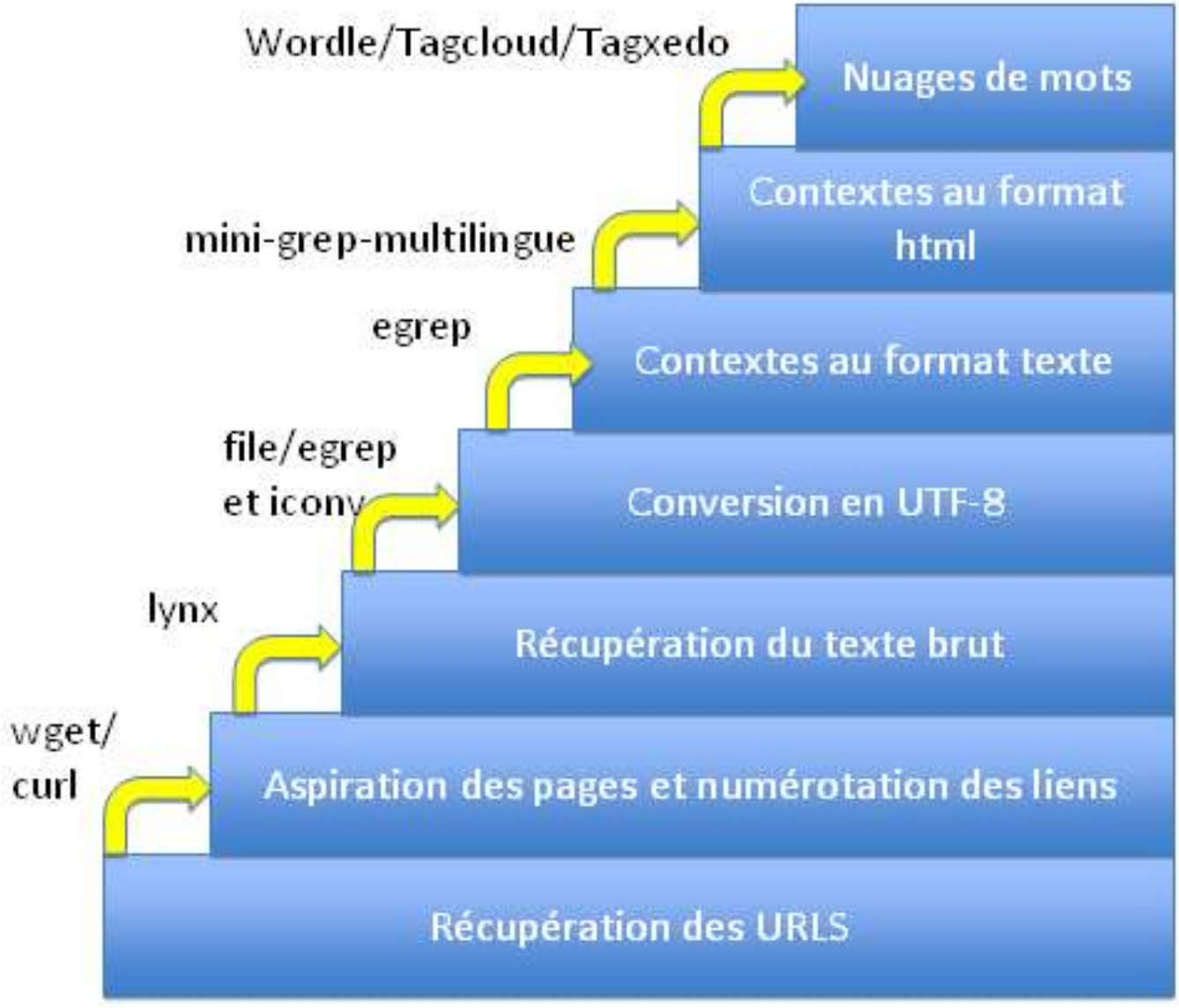

Un mot (et ses « variations » dans les langues choisies)
+ 1 hypothèse (?)
Quel type de données ?
Homogénéité des données ?
Phase 1 : construction du corpus
Une chaîne de traitements informatiques
Un corpus multilingue
Phase 2 : exploration du corpus
Sémantique lexicale
Approche textométrique
Outils « maison » :
iTrameur / le Trameur
Analyse

les nuages !!!

en allant vers les nuages

Exemples de projets passés


"un mot"
sur le web
sélection
50 URLs
(dans un fichier)
DOC
WGET
Option -i
pour traiter
le fichier d'URL
Pages
aspirées
stockées
localement
LYNX
Option -dump
pour ne garder
que le texte
DOC
"Fichier HTML"
vers
"Fichier texte"
EGREP
DOC
Filtrage
du texte :
Contexte de barrage
Organisation
Des données
PROJET-MASTER
Fichier Edition Affichage Favoris Outils ?
Précédente Rechercher Dossiers
Adresse C:\PROJET-MASTER
OK
PROJET-MASTER
Sélectionnez un élément pour obtenir une description.
Voir aussi :
Nom Taille Type
DOCUMENTATION Dossier de fichiers
DUMP-TEXT Dossier de fichiers
JOURNAL Dossier de fichiers
LISTES-URL Dossier de fichiers
PAGES-ASPIREES Dossier de fichiers
CONTEXTES Dossier de fichiers



Site de Huiyun WU (Inalco), Uyên-To DOAN-RABIER (Inalco), Mohamed Sofiane KERROUA (Inalco)

La structure d'un projet

Le projet : ça n'est que quelques ligne de code finalement

Étant donné un fichier d'URL (le programme pourra traiter plusieurs fichiers d'URLS...)

- Pour chaque URL ⟸ Boucle : for
  - Récupérer la page ⟸ Commandes : curl, wget
  - la stocker localement dans le dossier idoine
  - Si pas d'erreur lors de la récupération ⟸ Condition : if
  - Alors
    - Si la page est en UTF-8
    - Alors
      - En extraire le texte ⟸ Commande : lynx
      - En extraire des contextes autour des mots choisis ⟸ Commande : egrep
    - Sinon ⟸ Alternative à if : else
      - On essaie de détecter l'encodage de la page ⟸ Commande : file
      - Si l'encodage est reconnu
      - Alors
        - Extraire le texte ⟸ Commande : iconv
        - Conversion en UTF-8
        - En extraire des contextes autour des mots choisis
      - Sinon
        - On ne fait rien
  - Sinon
    - On ne fait rien
  - Une URL traitée = création dans un tableau HTML d'une ligne où on stocke les éléments créés :
    - Liens vers page, page avec contenu textuel (utf-8 et encodage initial), contexte autour du pôle

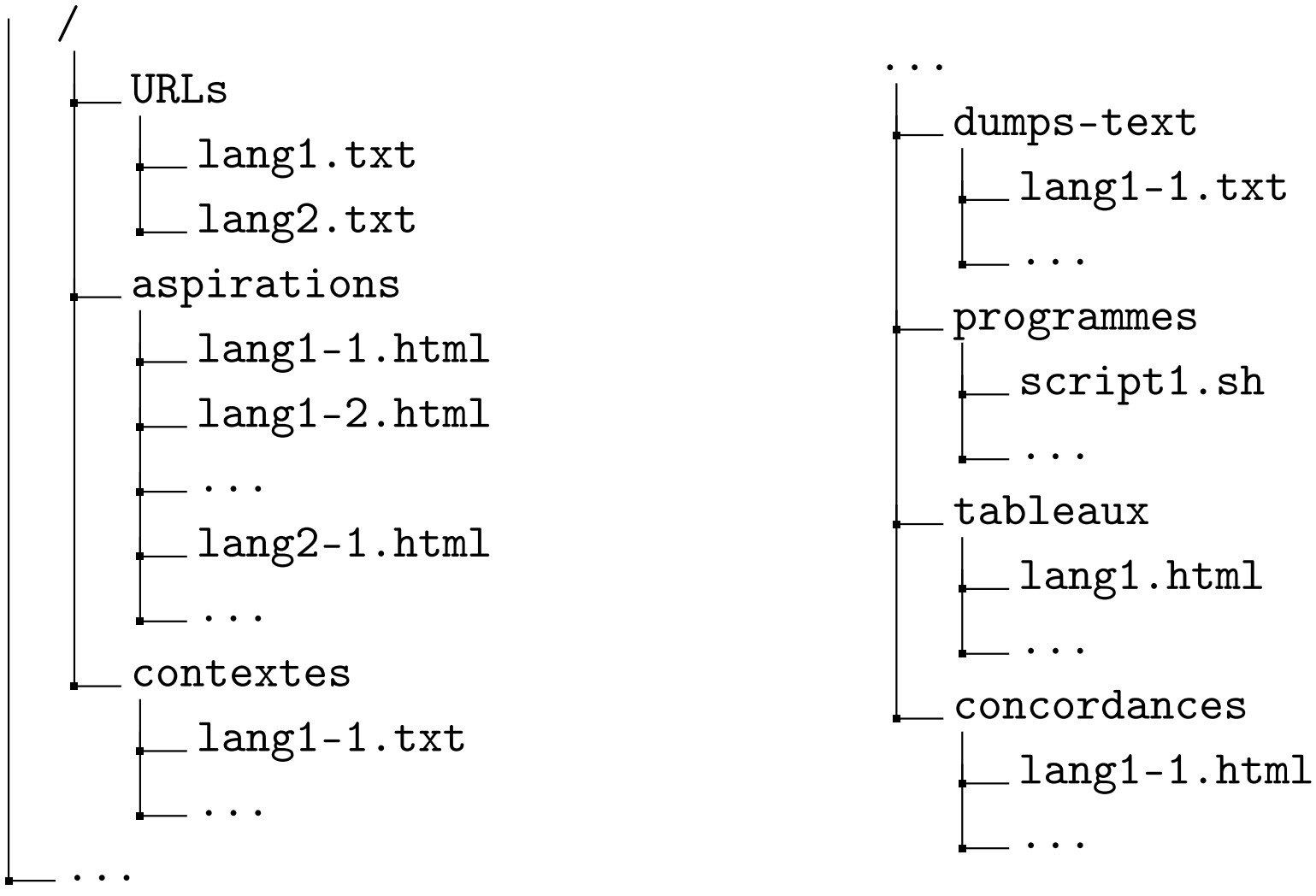
Vous n'avez plus qu'à construire la clé...

On va vous aider !

Étapes

- Toutes les étapes sont décrites en ligne
- Phase 1 :
  - Les différentes étapes d'écriture des scripts de traitement des pages contenant les mots choisis
  - http://www.tal.univ-paris3.fr/cours/PROJET-MOT-SUR-LE-WEB/
- Phase 2 :
  - Des nuages et des arbres de mots "multilingues" sur le web
  - http://www.tal.univ-paris3.fr/cours/PROJET-MOT-SUR-LE-WEB/nuages.html

Structure dossier projet



Contenu du tableau (pour chaque langue)

Dans les tableaux, on attend, *a minima*, es colonnes suivantes :

1. numéro
2. URL
3. code
4. encodage
5. nombre d'occurrences du mot dans la page
6. page HTML brute
7. dump textuel
8. concordancier HTML

Pour avoir tous les points, il faudra aussi rajouter d'autres colonnes : bigrammes, gestion des fichiers robots.txt, concordancier avec coloration des mots spécifiques dans les contextes.

La suite

Lancer les scripts `PALS` pour avoir les mots qui gravitent autour du mot de votre choix. Ces cooccurrences seront à la base de vos analyses.

On pourra améliorer le concordancier en mettant une emphase sur les mots spécifiques dans les contextes (couleur, surlignage, etc) pour un retour au texte plus efficace.

Il faudra également créer des nuages de mots avec le programme `wordcloud`.

On fera alors une analyse par langue et une conclusion globale.