De l'ordre dans les dictionnaires au GETA

Introduction

Au fil des discussions, il apparaît clairement que le GETA a besoin de mettre un peu d'ordre dans ses dictionnaires. Il faut rassembler tous les dictionnaires UNL, récupérer les dictionnaires Ariane, reprendre le stock de dictionnaires de Haï. Pour l'instant, les dictionnaires sont éparpillés tant et si bien que l'on ne peut pas vraiment faire un inventaire des ressources.

Description

Je propose de répertorier toutes les ressources dictionnairiques du GETA et de les installer dans une seule hiérarchie régie par un ensemble de règles. Une seule personne sera responsable de cet entrepôt de dictionnaires (moi dans un premier temps).

Les utilisateurs pourront :

•consulter certaines ressources à l'aide de dicoweb,

Deux types d'utilisateurs sont envisageables :

- les membres du CLIPS qui peuvent consulter toutes les ressources accessibles par dicoweb ;

- le public qui pourra consulter les ressources qui seront accessibles au public répertoriées pour le projet SILFIDE.

•chercher une ressource dont ils ont besoin.

Un module de recherche sera bientôt mis en place. Parmi les critères de sélection, la langue source, les langues cibles, la date de création seront disponibles.

•ajouter ou modifier une ressource existante selon les droits qui lui sont octroyés.

Dans un premier temps, il s'agit seulement de classifier les ressources en les regroupant dans un même endroit. Par la suite, des outils de manipulation de ces ressources pourront être développés.

Implémentation

Avec Marie-Hélène Corréard, nous avons achevé de mettre en place au centre de recherches de Xerox une structure similaire.

La machine

Pour réutiliser au maximum les outils présents et futurs, je propose d'installer un hiérarchie de fichiers et répertoires sur une machine UNIX du laboratoire.

Trois contraintes conditionnent le choix de la machine :

• Perl doit être installé sur la machine,

• Elle doit faire tourner un serveur web pour utiliser dicoweb,

• La hiérarchie doit être sauvegardée régulièrement.

Je propose d'utiliser Matahari si Perl est installé et si il est possible de monter l'arborescence de matahari sur isis pour pouvoir accéder aux dictionnaires avec dicoweb.

Sinon, il serait plus facile d'utiliser isis. Deux problèmes demeurent :

•cette machine n'est pas une machine du GETA,

•Y a-t-il suffisament de place sur isis pour tous les dictionnaires du GETA ?

La hiérarchie

Quelques règles simples seront utilisées pour construire la hiérarchie :

•Tous les dictionnaires seront situés dans /Dict,

•Les dictionnaires encodés en XML seront situés dans /Dict/XML/

•Chaque dictionnaire est dans un répersoire dont le nom est composé comme suit : /Dict/DirName_la-lb-lc/ où
- DirName est le nom du dictionnaire par exemple FeM. Il commence par une majuscule. la, lb, lc sont les langues contenues dans le dictionnaire classées par ordre alphabétique. Le nom des langues suivra la norme ISO-639, par exemple en, fr. (voir plus loin).

Par exemple /Dict/FeM_en-fr-ml/ pour le FeM source.

/Dict/XML/FeM_en-fr-ml/ pour le FeM encodé en XML.

•Chaque répertoire de dictionnaire contiendra un README qui donnera les informations suivantes :

```
------ information générale(sur le dictionnaire) ---
nom: ex Français-anglais-Malais
categorie: ex trilingue
langue source:
langues cible:
contenu:
domaine: ex general
date de création:
auteurs:
responsible:
info:
---------- liste des fichiers -------------
```

pour chaque fichier:

nom:

date de creation:

auteur:

version:

encodage:

info:...

L'information spécifique au contenu d'un dictionnaire sera écrite dans un fichier dont le nom sera ajouté à la liste des fichiers du répertoire.

•Chaque fichier contenant du texte d'un dictionnaire sera composé comme suit :

dictName_sl_tl1_tl2-encoding-version.extension où

- dictName est le nom du dictionnaire par exemple fem

- sl est la langue source du dictionnaire

- tl1, tl2 sont les langues cibles du dictionnairetriées par ordre alphabétique. Le nom des langue doit suivre la norme ISO-639

- encoding est le type d'encodage du fichier (ISO-8859-[1-9], ASCII, UTF-8, SJIS, etc). Si le fichier est en XML, ce renseignement n'est pas nécessaire. De plus, si l'encodage n'est pas indiqué, c'est ISO-8859-1 par défaut.

•- version est le numéro de version du fichier. Ce numéro est optionnel

- extension est l'extension du fichier (txt, rtf, xml, html, sgml, etc.)

Par exemple, /Dict/FeM_en-fr-ml/fem_en_fr-macroman-v1.txt

/Dict/XML/FeM_en-fr-ml/fem_en_fr-v1.xml

Pour l'instant, tous les fichiers et dossiers seront accessibles en lecture et écriture par le responsible et seulement en lecture par le groupe. Par la suite, des évolutions seront nécessaires.

ISO 639 1988 :

aa|ab|af|am|ar|as|ay|az|ba|be|bg|bh|bi|
bn|bo|br|ca|co|cs|cy|da|de|dz|el|en|eo|
es|et|eu|fa|fi|fj|fo|fr|fy|ga|gd|gl|gn|
gu|ha|hi|hr|hu|hy|ia|ie|ik|in|is|it|iw|
ja|ji|jw|ka|kk|kl|km|kn|ko|ks|ku|ky|la|
ln|lo|lt|lv|mg|mi|mk|ml|mn|mo|mr|ms|mt|
my|na|nb|ne|nl|nn|no|oc|om|or|pa|pl|ps|
pt|qu|rm|rn|ro|ru|rw|sa|sd|sg|sh|si|sk|
sl|sm|sn|so|sq|sr|ss|st|su|sv|sw|ta|te|
tg|th|ti|tk|tl|tn|to|tr|ts|tt|tw|uk|ur|
uz|vi|vo|wo|xh|yo|zh|zu

Voir http://indy.culture.fr/SemUnicode/langues.htm pour plus d'informations.

Annexe : DTD XML des README

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<!DOCTYPE dictionary-readme [
<!ELEMENT dictionary-readme (readme-info,general-info,files-
list)>
<!ELEMENT readme-info (creation-date, author)>
<!ELEMENT general-info (name, category, source-language,
target-language*, domai
n, creation-date, author+, responsible+, info?)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT category (#PCDATA)>
<!ELEMENT source-language (#PCDATA)>
<!ELEMENT target-language (#PCDATA)>
<!ELEMENT domain (#PCDATA)>
<!ELEMENT creation-date (#PCDATA)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT responsible (#PCDATA)>
<!ELEMENT info (#PCDATA)>
<!ELEMENT files-list (file+)>
<!ELEMENT file (name,creation-
date,author+,encoding?,version?,info?)>
<!ELEMENT encoding (#PCDATA)>
<!ELEMENT version (#PCDATA)>
]>
```