Voici le compte-rendu de la réunion du projet innovant du 18 janvier 2005

Etaient présents Frédérique Bénard, Serge Fleury, Cédric Gendrot, Michel Jacobson, Thierry Pagnier, Patrick Renaud, Pollet Samvelian.

##
DATE DE PROCHAINE REUNION ….
Nous avons décidé de la fixer au lundi 14 février à 9h30.
J'espère que cette date conviendra également à ceux qui n'étaient pas présents. N'hésitez pas à me prévenir si ce n'est pas le cas.
##

Voici les points principaux qui ont été abordés :

1- Description de nos corpus …

Suite à notre précédente réunion, Frédérique (avec l'aide de Michel) nous a fait un compte-rendu de la dernière version d' OLAC (1.0). Il semble que les anciennes versions soient abondamment mélangées avec les plus récentes et que l'on puisse avoir accès à toutes ces versions sans savoir très facilement de quelle version il s'agit. C'est le cas pour les articles que nous avons en notre possession (version 0.4). De même pour le livre (version 0.4) qui a circulé lors de cette dernière réunion (« Meta Data in practice. » D. Hillman et E. Westbrooks). En cherchant (beaucoup) sur le site d'OLAC, il est possible de trouver cette dernière version 1.0.

Je rappelle qu'OLAC ajoute des attributs aux éléments de DC afin de les compléter pour une meilleure description des corpus linguistiques. A présent, la version d'OLAC 1.0 propose les 5 attributs suivants (fourni par Frédérique) :

- discourse-type : qui s'applique à **type** et **subject** du DC.
- language (pour language identification) :qui s'applique à **language** et subject
- linguistic-field : qui s'applique au subject
- linguistic-type : pour type
- role : contributor (et creator)

La Synthèse « finale » de Dublin-Core et de la dernière version d'OLAC sera fournie prochainement par Frédérique.... Cette liste de descripteurs nous permettra de fournir une description (méta-données) des corpus que nous avons sur le serveur.
Quelques problèmes sont posés par cette description ; sans entrer dans le détail, il n'est pas toujours facile de savoir comment les attributs proposés par OLAC peuvent être remplis, mais il en va de même pour les éléments et raffinements de DUBLIN CORE. Il est donc plus difficile qu'il n'y parait de remplir cette feuille de description. Par exemple, il ne semble pas possible de laisser un néophyte remplir tout seul ces champs. Il nous sera nécessaire d'écrire des conventions pour conserver une forte cohérence dans nos descriptions de corpus.

Nous préparerons donc à 2 (Frédérique et moi-même) la description d'un de nos corpus (corpus aupelf urelf ... encore lui !!) et vous la présenterons … pour que chacun donne son avis, que l'on puisse se rendre compte des problèmes mentionnés ci-dessus, et que l'on discute des conventions à adopter.
L'étape suivante consistera à décrire (au moins) un corpus chacun et à confronter les résultats … je rappelle que le but de notre projet est de pouvoir décrire n'importe quel corpus qui nous sera fourni par la suite…. et ce de manière assez standardisée ...

2- Réorganisation des données ….

Comme mentionné dans le précédent compte-rendu : « Du type de corpus dépend un certain nombre d'habitudes ... certains travaillent avec des corpus longs qui ne peuvent être découpés ; d'autres utilisent des séries d'énoncés découpés en d'innombrables petits extraits . »
Un regroupement des fichiers doit être effectué pour le cas où une série d'extraits ne nécessite pas de feuilles de description bien distinctes … J'ai passé en revue les corpus que nous avons recueillis, ce regroupement peut être

fonction de l'objectif linguistique (ce qui est pertinent dans le corpus), mais peut dépendre de choix plus subjectifs. Il s'agit de toute façon d'un compromis, puisqu'un regroupement trop grossier ferait perdre de l'information, alors qu'un regroupement trop minutieux noierait l'utilisateur dans un surplus d'information.

Cependant, il devrait être possible d'accéder à la description du contenu (Je ne parle plus de la description des corpus/méta-données, mais bien la description du contenu /annotations) dans une requête au sein du serveur (par exemple uniquement le locuteur X, ou les phrases Y dans tel corpus).

Affaire à suivre !!!

3- Présentation de Kepler

Une fois que la feuille de description aura été finalisée et que les principes (conventions) pour remplir cette dernière auront été définis, notre objectif consistera à proposer un petit formulaire/outil qui nous permettra de remplir les différents champs de façon ergonomique.
Serge nous a donc présenté KEPLER, un outil qui permet d'insérer les éléments de Dublin Core et D'OLAC dans une feuille de description, au moyen de différentes boites de dialogue.

Ce type d'outil doit intégrer :
- une interface : ensemble de méta données à insérer via un formulaire .php
- une base de données MySql et un serveur Apache pour enregistrer et afficher les méta données insérées.

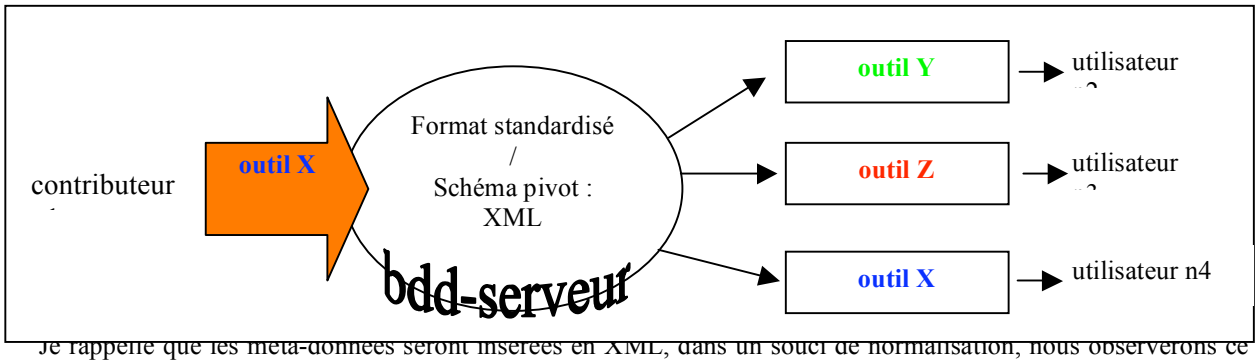
Puisque KEPLER semble très compliqué à utiliser et apporte autant de problèmes qu'il ne saurait en résoudre, Serge se propose de programmer ce genre d'outil ... si j'ai bien compris, il ressemblera beaucoup à ce qui se trouve déjà sur le serveur ... je vous rappelle le lien :
http://pi-ed268.univ-paris3.fr/files/bddpied268-query_metadata/bddpied268.php

4- Présentation de quelques logiciels utilisables pour l'annotation de corpus.

Rappel : Nous avions discuté lors de la précédente réunion de l'utilité de montrer aux différents membres de l'ED (susceptibles de nous fournir des corpus) quels types d'outils ils pouvaient utiliser pour annoter leurs corpus, afin que ce soit un travail plus simple, rapide et efficace pour eux (et pour nous également). L'objectif principal étant que plus aucun étudiant n'annote son corpus sous Word, et en perdant donc toute l'information temporelle qui nous est si chère.

Luiggi n'ayant pu se libérer, nous n'avions pas tous les éléments pour faire une petite descriptions de ces outils .... Ce qui ne nous a pas empêché d'en discuter longuement :

comme toujours, l'aspect « pluridisciplinarité » qui nous est si cher nous amène bon nombre de difficultés, ... Puisque enfermés dans nos habitudes concernant l'annotation de corpus, nous ne nous rendons pas compte de la grande variété des annotations ... (N.B. Une nouvelle fois, je ne parle plus de méta-données mais plutôt des annotations du contenu …)
Il faut que les corpus puissent être stockés dans un format standardisé (non propriétaire), format qui serait un schéma pivot comme selon la figure suivante



Je rappelle que les méta-données seront insérées en XML, dans un souci de normalisation, nous observerons ce même principe pour l'annotation : l'outil utilisé pour l'annotation de corpus devra donc être compatible avec XML. Le problème principal est que chaque discipline utilise un petit nombre d'outils très pratiques pour

certains aspects précis, donc très utilisés au sein d'une discipline, mais qui ne sont pas forcément très compatibles .... Malgré tout des passerelles peuvent être construites pour passer du format propriétaire fourni par certains logiciels au format standardisé qu'est XML. Certaines existent déjà !

Ce n'est malheureusement pas le seul problème ...Par exemple nous devons faire en sorte que les caractères utilisés pour tel corpus puissent être lus par n'importe qui ! Si je ne m'abuse, pour parer à ce genre de problèmes, le site du LACITO fournit les outils nécessaires, ou bien propose une visualisation en ligne. Par exemple, UNICODE a été créé pour résoudre ce genre de problèmes, mais tous les outils n'intègrent pas UNICODE (Cf. Praat)

Comme prévu dans le programme de la réunion, nous en sommes venus tout naturellement à la TEI (Text Encoding Initiative) au travers de cette discussion sur la standardisation de l'annotation. Voici un rappel sur la TEI ... (pris sur le site web qui lui est dévoué : http://www.tei-c.org (http://www.tei-c.org/ns/1.0)

"They [the TEI guidelines] provide means of representing those features of a text which need to be identified explicitly in order to facilitate processing of the text by computer programs. In particular, they specify a set of markers (or tags) which may be inserted in the electronic representation of the text, in order to mark the text structure and other textual features of interest. Without such explicit markers, many important features remain difficult to locate by mechanical means such as computer programs, and thus difficult to process effectively. The process of inserting such explicit markers for implicit textual features is often called `markup', `encoding', or `tagging', and the term encoding scheme or markup language denotes the rules which govern the use of markup in a set of encodings.
The Guidelines formulated in this document are intended for use in interchange between individuals and research groups using different programs and computer systems over a broad range of applications. Since they contain an inventory of the features most often found useful for text processing, the Guidelines also provide help to those creating texts in electronic form. They can also be used for the local storage of text which is to be processed with multiple software packages requiring different input formats.

The Guidelines apply to texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content. They treat both continuous materials (`running text') and discontinuous materials such as dictionaries and linguistic corpora."

Cette partie se trouve aux frontières de notre projet puisque ne nous comptons pas, dans le cadre du projet, proposer une manière standardisée d'annoter les corpus. Cependant, il serait préférable de proposer des outils qui puissent intégrer à l'avenir ces annotations standardisées.
Ceci pose problème à la fois en amont et en aval de la base de données :

- il ne faut pas que les utilisateurs potentiels soient contraints par un outil qui leur semble peu pratique ... sinon ils ne l'utiliseront pas ... après tout, la principale motivation des étudiants est de faire un corpus simplement pour faire leur travail, et répondre à leurs hypothèses ...
- il faut que les corpus stockés sur le serveur puissent être analysés par n'importe quel utilisateur, avec des (ses ?) outils d'analyse. On en revient donc au problème mentionné ci-dessus : les outils d'analyse utilisés par chacun au sein d'une discipline ne sont pas nécessairement compatibles avec le format standardisé que nous utiliserons. Malgré tout des passerelles peuvent/doivent être construites pour passer du format standardisé qu'est XML au format propriétaire utilisé par certains logiciels. Certaines existent déjà, je pense à Praat (mais aussi CLAN, Transcriber est déjà basé sur XML) comme chacun sait, puisque je prêche pour ma paroisse, mais il faut penser à tout le monde, n'hésitez pas à mentionner des outils qui vous sont chers, et que je ne connais certainement pas!

Affaire à suivre !!!

5 – Programme pour la prochaine réunion

- Présentation par Frédérique et moi-même de la description du corpus Aupelf-Urelf en fonction de la feuille de description « finale » intégrant Dublin Core et OLAC. Nous vous transmettrons avant cette prochaine réunion une description préliminaire du corpus, ainsi que la feuille de description que nous utiliserons ... afin que vous puissiez réagir
- élargissement de cette description aux autres corpus ... chacun d'entre nous pourra mentionner les problèmes qui pourront survenir en fonction du (des) corpus qu'il ou elle a fourni(s).
- Suite de la discussion sur les logiciels/outils à utiliser pour l'annotation des corpus.

- Si tout va bien ... je montre l'enregistreur MARANTZ tant attendu (il est arrivé ...) !