ASSISES DE LA RECHERCHE
Université Sorbonne Nouvelle (Paris 3)
2 et 3 mars 2006 - Paris

“ Propositions de Normalisation pour une Base de Corpus Multimédia ”

à l’ED268

Serge Fleury, Cédric Gendrot, Michel Jacobson

Frédérique Bénard, Sonia Branca, Maria Candéa, Thierry Pagnier, Patrick Renaud, André Salem, Pollet Samvelian, Luiggi Samsonetti, Jacqueline Vaissière.

Une base de données

par et pour

L’École Doctorale 268

- Une approche pluridisciplinaire, et dont l'objectif évident est une contribution à la linguistique de corpus.

- **Parmi les membres du projet :**

  - Sociolinguistique : P. Renaud; S. Branca; T. Pagnier
  - Phonétique : C. Gendrot; M. Candéa; J. Vaissière
  - TAL / Informatique : S. Fleury; M. Jacobson; A. Salem ; F. Bénard
  - Syntaxe, Sémantique : P. Samvelian
  - Acquisition : L. Samsonetti

Utilités d'une base de corpus

- Archivage de langues. (LACITO :http://lacito.vjf.cnrs.fr/archivage/)
- Analyse / synthèse de la parole. (ELDA / LDC)
- Favoriser les collaborations pluri-disciplinaires. (sur un même corpus)
- Aider les jeunes chercheurs pour la constitution de leurs corpus (but pédagogique – stratégie cumulative)

Guide des bonnes pratiques

… pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux.

http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm

- **Séminaire BNF C-Oraux** 17 mai 2005

Guide des bonnes pratiques …

1- Recueil des données

2- Aspects juridiques

3- Transcription/Annotation

4- Exploitation, conservation, diffusion :

Notre travail !!

Le cœur du projet

1 L'élaboration **d'une base de corpus** (principalement oraux) regroupant des données de langues de différentes natures

→ Approche pluridisciplinaire

2 Des **propositions de normalisation** pour l'encodage de corpus de langue.

Exploitation, conservation, diffusion

- Nécessité de normaliser :
  - **pourquoi ?**

    Besoins de partager, diffuser et rechercher/retrouver

  - **Comment ? : Grâce à des métadonnées**

    XML → Dublin Core / OLAC

eXtensible Markup Language

Open Language Archive Community

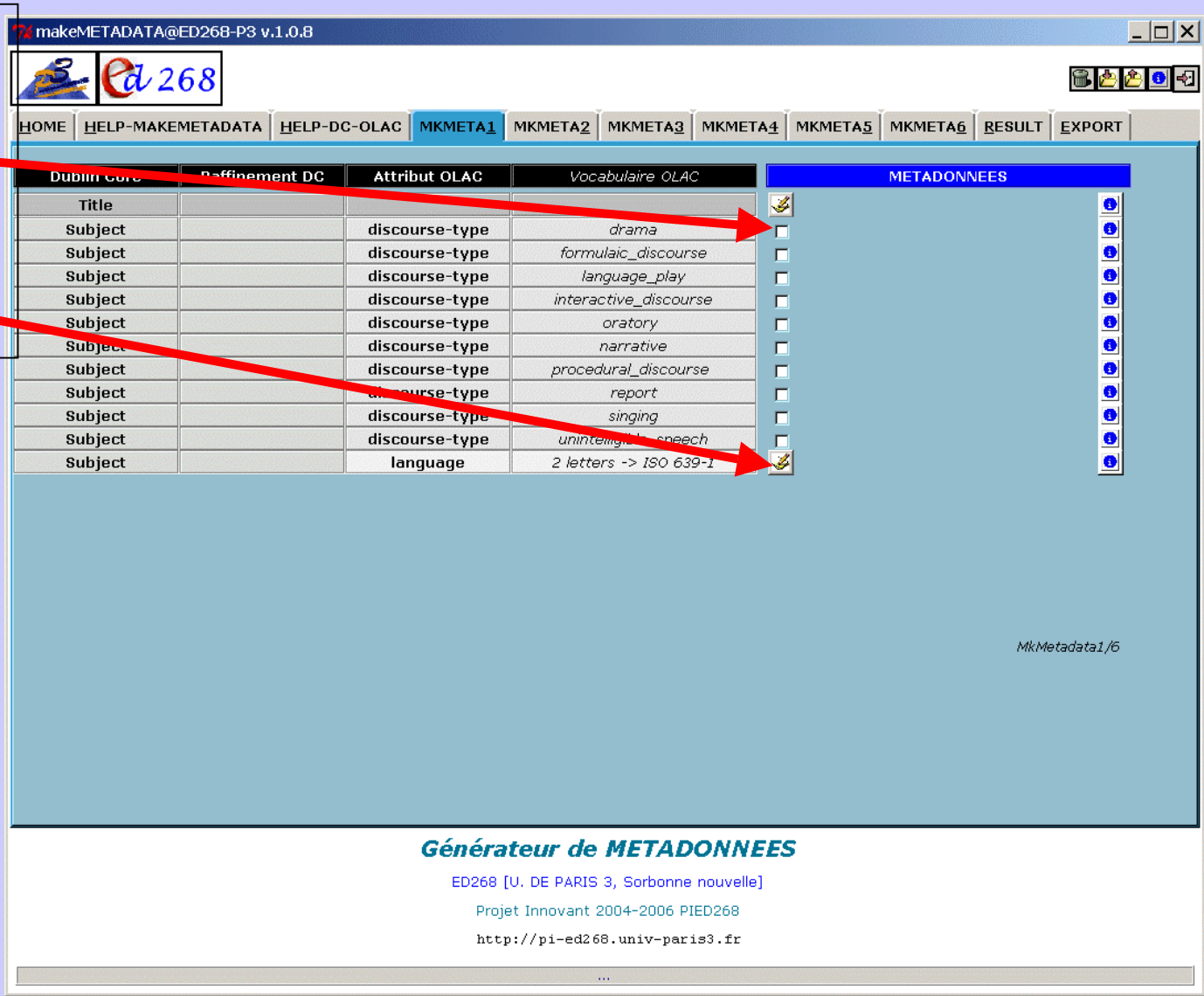
Résultats visibles …

Un Générateur de métadonnées : écrire de façon standardisée les données qui décrivent les données (ici les corpus)

il suffit de :

- cocher les cases

- compléter les espaces vides



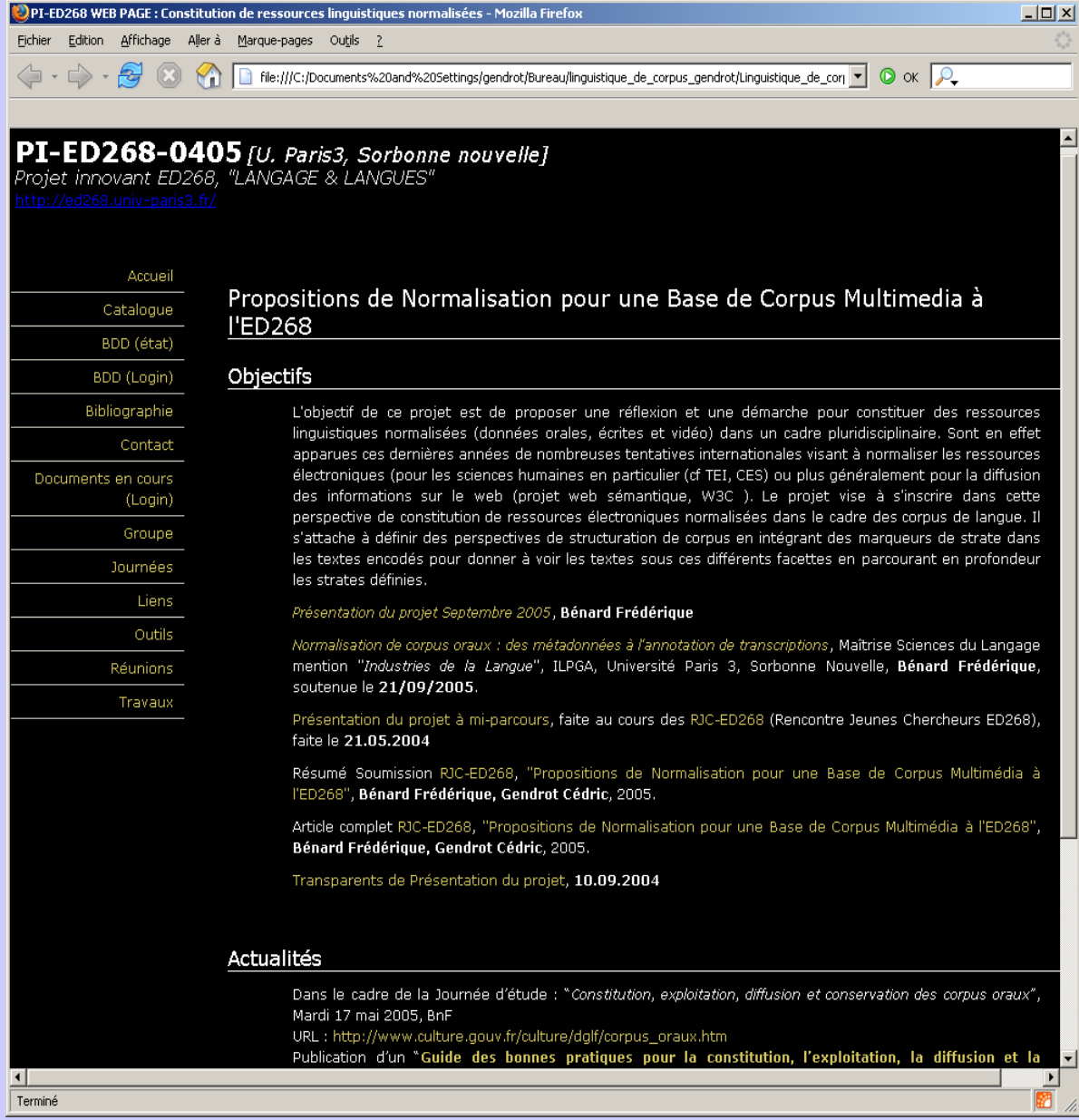
un site Web avec :

http://pi-ed268.univ-paris3.fr/

- les travaux/outils
- les présentations officielles
- des liens

et bien sûr …

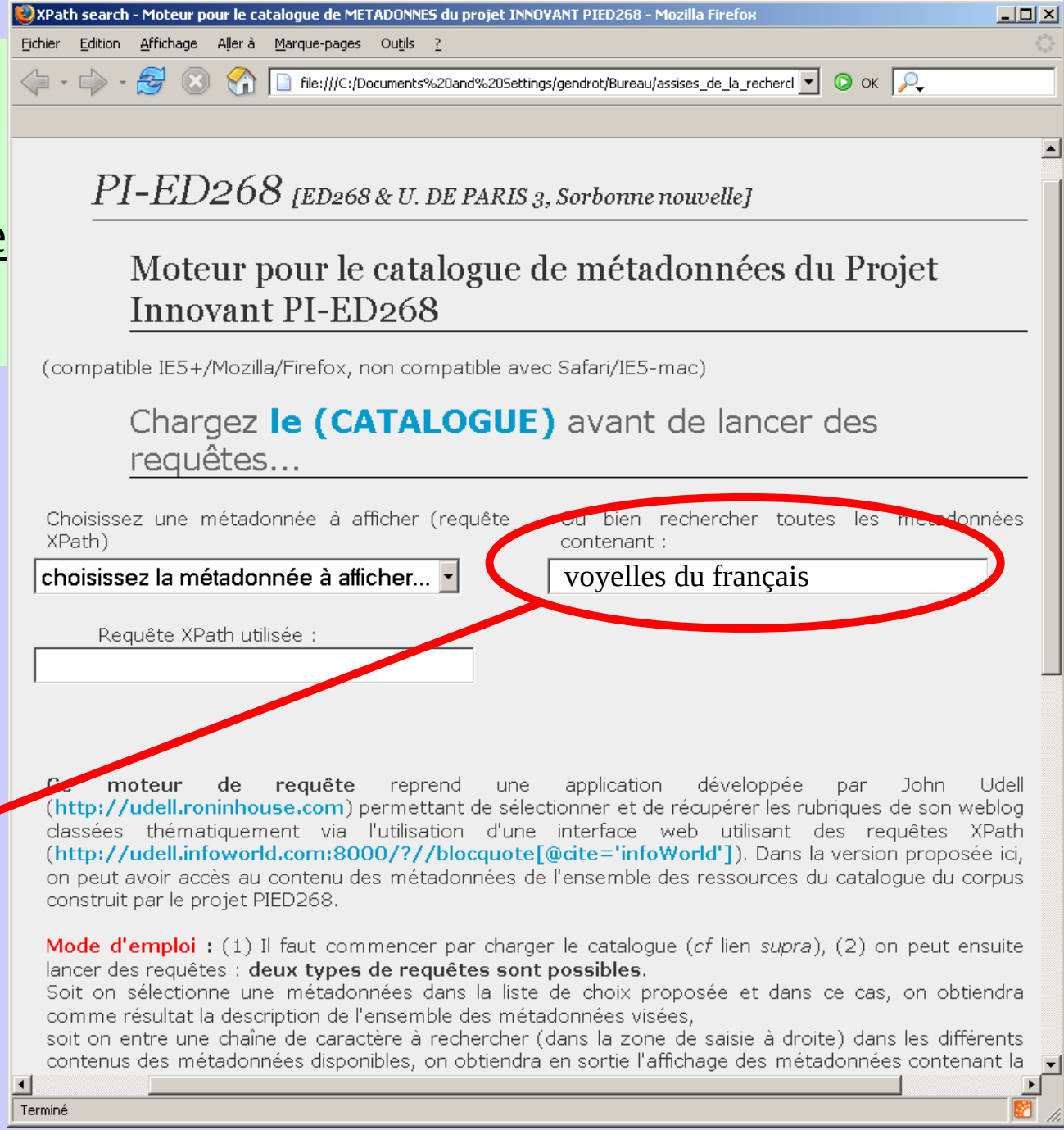
les Données et le MOTEUR de Recherche



et bien sûr …

les Données et le MOTEUR de Recherche

Taper un mot-clé pour rechercher des corpus



En résumé :

- Offrir aux membres de l'ED268 la possibilité de :
  - Mettre à disposition leurs corpus.
  - Utiliser pour leurs recherches les données stockées dans cette base de données

en cours …

- Amélioration du moteur de recherche …
- Candidature acceptée aux centres de ressources numériques proposés par le C.N.R.S.

dans un futur proche …

- **<u>extraire et/ou afficher des occurrences prononcées ...</u>**

  cela nécessite :

  - un moteur de recherche qui travaille sur l'annotation et non les métadonnées (facile)
  - que les annotations soient complètes et compatibles (+ difficile)

Merci de votre attention

Contacts et Liens …

- Site Web :

http://pi-ed268.univ-paris3.fr/

- Serge Fleury : serge.fleury@univ-paris3.fr
- Cédric Gendrot: cgendrot@univ-paris3.fr