J-M. Eloy, F. Martin, C. Rey

LESCLAP (CERCLL)

Université de Picardie Jules Verne

fanny.martin@u-picardie.fr

jean-michel.eloy@u-picardie.fr

christophe.rey@u-picardie.fr

Atelier TALARE 2

22 juin 2015

Caen

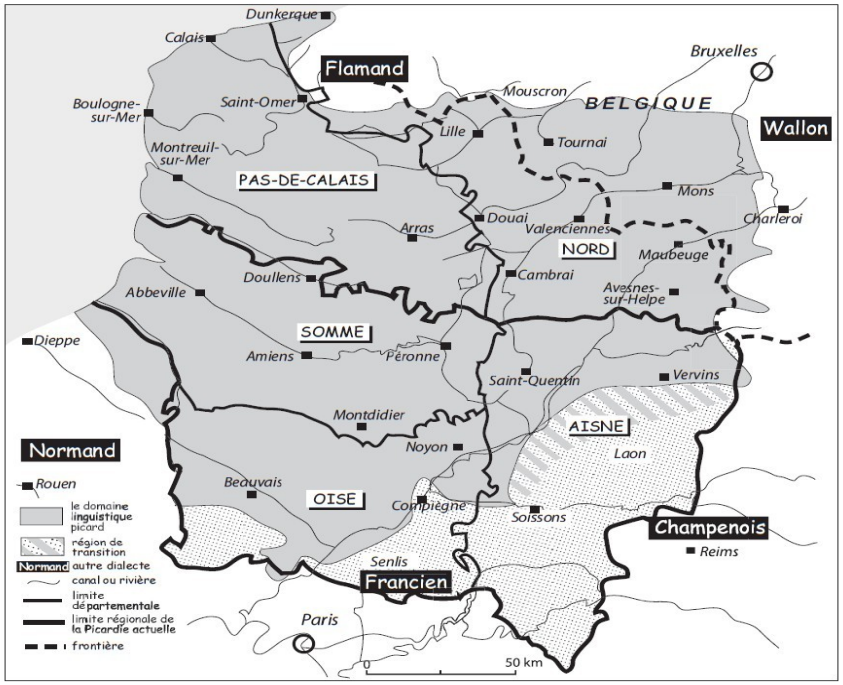
PICARTEXT : Une ressource informatisée pour la langue picarde

UNIVERSITÉ de Picardie Jules Verne

LESCLAP
Linguistique et sociolinguistique : contacts, lexique, appropriations, politiques

La langue picarde, langue de France

* Longue et riche histoire
* Activité de publication dense
* Domaine linguistique versus Région administrative
* Langue collatérale



Un développement sans standardisation

* Absence de norme centralisatrice

* Logique de pôles *("standardisation de par en bas vs standardisation de par en haut")=>autre forme d'aménagement linguistique* - Cette logique des pôles de pratique est-elle une chance ou une entrave à un développement homogénéisant de la langue ?

* Une extrême variation linguistique (orthographique, lexicale, syntaxique, etc.)

Exemple de variation graphique :
« c'était » *: ch'étoué, ch'étouait, ch'étouis, ch'étouo, ch'étwo, ch'étouot, ch'étwot,* et même *ché toué*, etc.

Quelques éléments de présentation de Picartext

- Un projet réalisé au sein du laboratoire amiénois LESCLAP (CERCLL-4283) et co-dirigé par Jean-Michel ELOY et Christophe REY
- Un projet de recherche réalisé grâce au soutien financier du Conseil Régional de Picardie – 2008-2011.
- Réalisation rendue possible par l'embauche de nombreux vacataires, d'étudiants et de 2 post-doctorants successifs (Yayoï Nakamura-Delloye et Alain Dawson).

- Une base textuelle comprenant environ dix millions de mots
- Une base de données littéraire composée d'éléments divers (dictionnaires, contes, recueils de poésies, romans, chansons, etc.)
- Une base panchronique (Textes depuis le XVIII^{e} jusqu'au XXI^{e} siècle)

UNIVERSITÉ de Picardie Jules Verne

LESCLAP Linguistique et sociolinguistique : contacts, Lexique, appropriations, politiques

CEP CENTRE D'ETUDES PICARDES Université de Picardie · Jules Verne

PICARDIE LA RÉGION

Picartext : pour quoi faire ?

1. Action Linguistique…

Langues collatérales « des variétés proches – objectivement et subjectivement –, aux plans linguistique, sociolinguistique et historique ou glottopolitique, les variétés tendanciellement en contraste étant historiquement liées par les modalités de leur développement ».

Faire émerger le commun - unitaire (koïnè) – élaborer la standardisation

Maîtriser les phénomènes de variation, notamment la variation graphique

Mise en place d'un **matériel pédagogique** : faciliter l'apprentissage de la langue

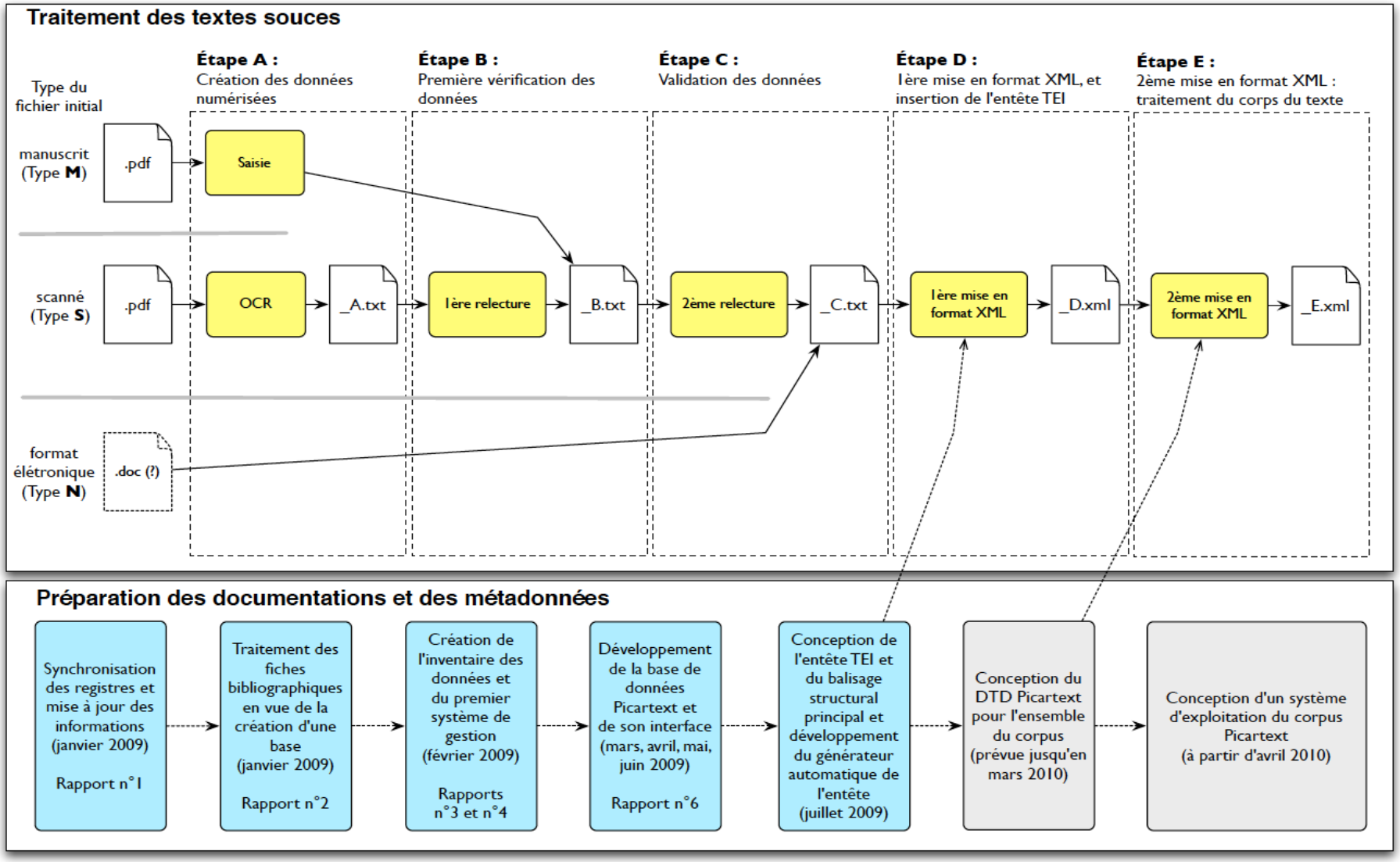
2. Action culturelle

Soutien à la **création littéraire**

Livraison de ressources textuelles (dictionnaires, oeuvres littéraires majeures, etc.)

La voie du Traitement Automatique des Langues

Aperçu rapide sur le processus de constitution de Picartext



Depuis 2010 : Mise en place d'un module d'interrogation de la base

PICARTEXT et les technologies du langage

- **http://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/index.php**

Une base de données MySQL en libre accès dotées d'environ 3,5 millions de mots interrogeables

Une ressource permettant de sonder l'ensemble du vaste domaine linguistique picard selon plusieurs méthodes de recherche :

Maîtrise de la variation

* Chaîne littérale : le mot est recherché sous la forme exacte fournie par l'utilisateur
* Correspondance phonétique: le mot est recherché sous les différentes formes orthographiques utilisées par les auteurs, à condition que la prononciation soit identique
* Correspondance dialectale : le mot est recherché sous toutes ses formes théoriquement possibles en picard, y compris avec d'autres prononciations que celle qui est fournie
* Expression rationnelle étendue : recherche grâce à des expressions régulières

q Recherches affinées :

* Sélection d'un empan temporel relatif aux dates de naissance des auteurs
* Sélection d'une zone géographique de naissance des auteurs
* Sélection d'un genre textuel particulier

Les perspectives du projet RESTAURE

lilpa
linguistique, langues, parole
Alsacien

Clle ERSS
Occitan

LESCLAP
Picard

Traitement Automatique des Langues

Attentes et objectifs informatiques

Pour RESTAURE

- Mettre en place des outils de désambiguïsation grammaticale et lexicale, etc. (lemmatisation, étiquetages divers, etc.) qui supposent de maîtriser la variation ou les variations
- Parvenir à une véritable transversalité dans les 3 langues de travail du projet RESTAURE

Pour PICARTEXT

- Permettre d'étendre le nombre de textes informatisés en langue picarde
- Constituer un outil lexicographique de référence grâce aux nouvelles technologies (concaténation de répertoires déjà existants)

Pour le TAL

- Apporter in fine des éléments d'information permettant de faire évoluer les outils du TAL sur la variation dans les langues déjà richement dotées

Conclusions

Dans le contexte actuel, la langue picarde doit s'appuyer sur toutes ses richesses pour pouvoir figurer dans la liste des langues retenues dans le cadre de ratification de la Charte Européenne des Langues Régionales ou Minoritaires : et PICARTEXT en est une

Le projet RESTAURE constitue à la fois un prolongement possible pour PICARTEXT, mais aussi une occasion de dépasser cette première initiative

Le fait de pouvoir s'appuyer sur les outils du Traitement Automatique des Langues est une opportunité pour le picard de pouvoir mettre en place ce qui lui fait – peut-être – défaut aujourd'hui, à savoir une forme de standardisation ou d'homogénéisation qui pourrait être salutaire.

Nous ne perdons pas de vue que cette langue régionale, au même titre que celles également intégrées dans le projet RESTAURE, constitue également une opportunité de pouvoir faire progresser le traitement des « grandes » langues.