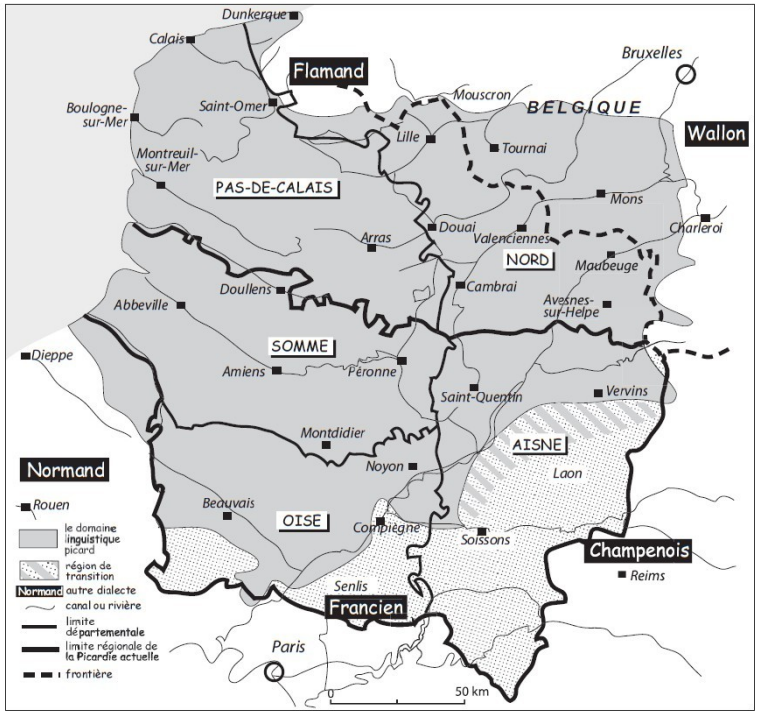
2e journée professionnelle du réseau Occitanica
CIRDOC – Béziers – 28 mai 2015

Le patrimoine numérisé et la recherche appliquée en TAL
PICARTEXT

https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/



1. Quelques éléments de présentation

- 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. (Projet blanc)
- Réalisation rendue possible par l'embauche de nombreux vacataires, d'étudiants et de 2 post-doctorants successifs **(moyens humains importants)**.

- Une base textuelle comprenant environ **dix millions de mots**
- Une base de données **littéraire** composée d'éléments divers (dictionnaires, contes, recueils de poésies, romans, nouvelles, chansons, etc.)
- Une base **panchronique et représentative de la variation du picard sur l'ensemble de son domaine linguistique**
- Textes depuis le XVIII^{e} jusqu'au XXI^{e} siècle
- Textes couvrant le territoire depuis l'Oise jusqu'au Hainaut belge.

2. Les objectifs de PICARTEXT

Action Linguistique…

Faire émerger le commun - unitaire (koïnè) – élaborer la standardisation (contexte patriculier d'une variation linguistique importante)

Élaboration de la norme graphique

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

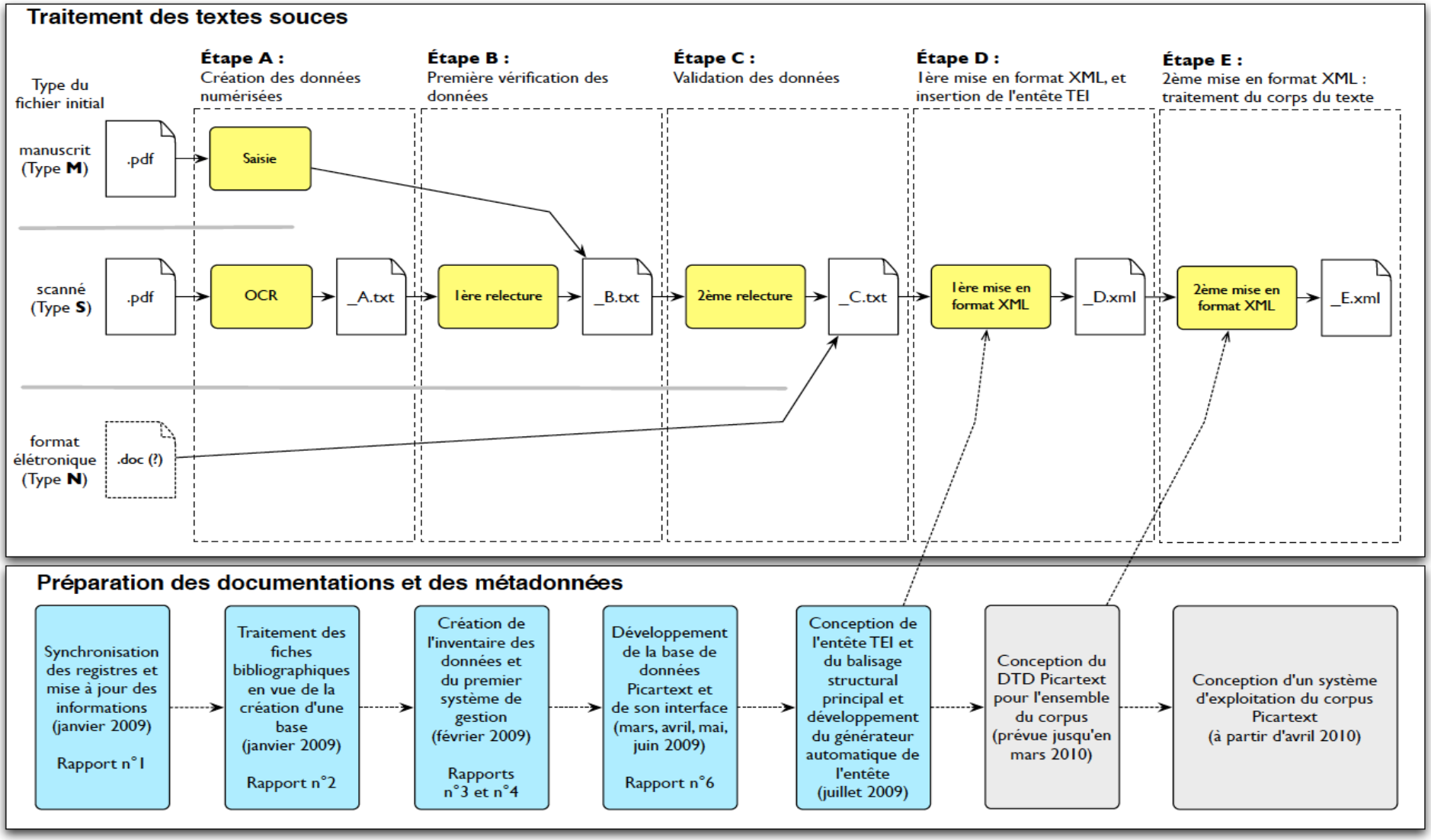
et culturelle

Soutien à la création littéraire, aux jeux…

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

Projet à la fois universitaire et culturel

3. Aperçu rapide sur le processus de constitution de Picartext



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

4. PICARTEXT et les technologies du langage (1)

- Une base de données MySQL en libre accès
- Environ 3,5 millions de mots interrogeables (pas la totalité)
- Une ressource permettant de sonder l'ensemble du vaste domaine linguistique picard selon **plusieurs méthodes de recherche** :
-

* 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 => **Neutralisation de la variation graphique (ex: éfant/éfan/effant)**

* 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 => **Neutralisation de la variation géographique (ex: kien/tchien/chien, etc.)**

* 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

5. PICARTEXT et les technologies du langage (2)

- Un projet désireux de **s'inscrire dans la galaxie des travaux utilisant des standards de balisage des données** (En-tête XML et à terme adaptation des DTD TEI pour baliser d'autres portions des textes.)
- Un **nouveau souffle** grâce à la participation du LESCLAP à l'ANR RESTAURE (RESsources informatisées et Traitement AUtomatique pour les langues REgionales):
- Bénéficier de la synergie d'équipes travaillant sur les langues régionales et possédant des expertises variées en TAL ((**LIMSI**) CNRS – UPR 3251, (**CLLE ERSS**) UMR 5263 CNRS & Université Toulouse 2, (**LiLPa**) EA 1339 Université de Strasbourg, **LESCLAP** (CERCLL – EA 4283)
- Complexifier les traitements informatiques déjà disponibles (lemmatisation, maîtriser la variation graphique, étiquetages divers, etc.)
- Valoriser un patrimoine en lui donnant une plus grande visibilité grâce aux apports du numérique.

Un effort important de valorisation est en train d'être mené pour faire en sorte que cette base devienne tout autant un objet pour les chercheurs que pour le grand public.