Offre d’emploi - projet EXO-POPP - Développeur Full-stack

Auteurs: Thomas CONSTUM, Pierrick TRANOUEZ, Thierry PAQUET
Affiliation : LITIS - Université de Rouen Normandie

Présentation du projet

Le projet EXO-POPP a pour but de développer une base de données de 300 000 actes de mariage de Paris et de sa banlieue entre 1880 et 1940. Ces certificats de mariage fournissent un grand nombre d’informations sur les mariés, leurs parents et les témoins de mariage.

Construire une telle base de données à la main nécessiterait au moins 50 000 heures de travail. Mais grâce aux récents développements en matière d'apprentissage profond, il est désormais possible de construire d'énormes bases de données avec des systèmes de lecture automatisés comportant la reconnaissance de l'écriture manuscrite et la reconnaissance d’entités nommées (extraction de mots-clefs).

Développement d’une plateforme web d’annotations de donnée

PIVAN

Afin d’appliquer des méthodes d’apprentissage profond sur des documents historiques, il est nécessaire d’annoter des données pour chacune des tâches du domaine (segmentation de documents, reconnaissance d’écriture, extraction d’entités nommées). Pour cela, nous utilisons au sein du LITIS la plateforme Pivan. Cette plateforme Web (front-end React, back-end Spring boot) permet de visualiser des données, ainsi que leurs annotations en segmentation, en reconnaissance d’écriture et en extraction d’entités nommées. Pivan permet également de renseigner les transcriptions des documents et d’annoter les entités nommées présentes dans le texte. Cependant, l’annotation du *layout* des documents n’est pas implémentée : indiquer quelles parties de l’image sont des paragraphes de texte, des titres, des illustrations etc. Cette fonctionnalité est en revanche disponible sur PIVAJ, un autre logiciel du LITIS permettant la visualisation et l’annotation de journaux.

Le rôle de ce poste sera dans un premier temps de s’inspirer du fonctionnement de PIVAJ pour ajouter une fonction d’annotation du *layout* à PIVAN, permettant ainsi de compléter la chaîne de traitement de la création de la vérité terrain initiale à la visualisation des résultats de productions finaux.

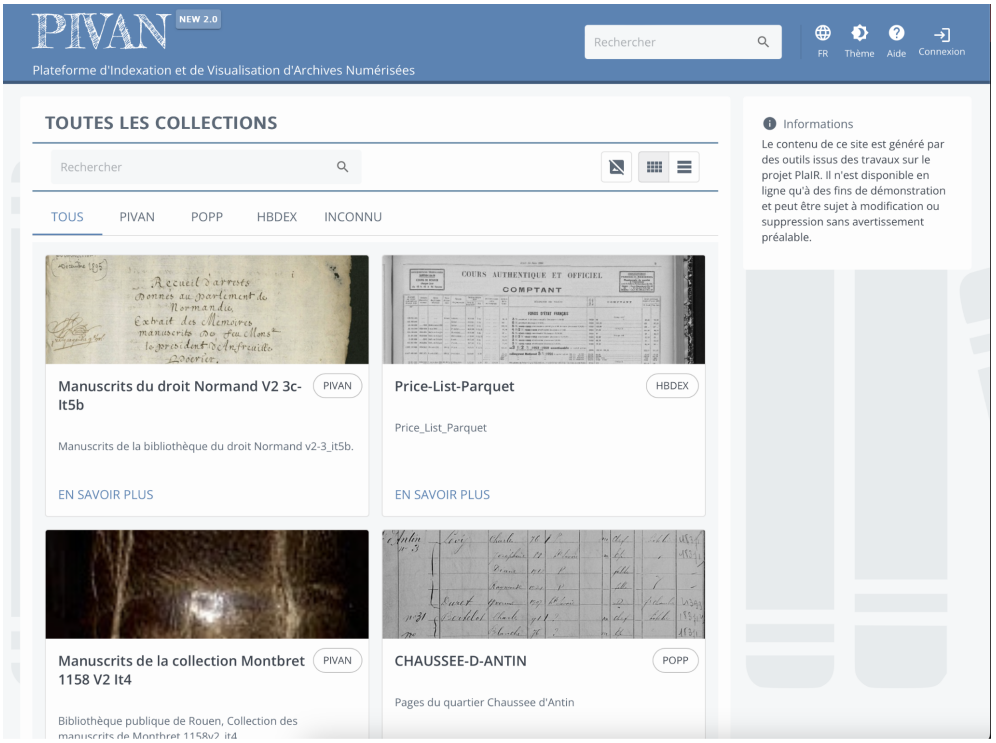

Figure 1 : Visualisation des collections de documents dans PIVAN
http://litis-pivan.univ-rouen.fr

Développements

Dans une première étape, pour prendre en main les développements existants, l'ingénieur(e) procédera à quelques corrections de bugs et améliorations mineures dans la gestion des collections.

Dans un second temps, il ou elle s'attaquera à l'annotation de layout.

En complément de cette tâche, d'autres fonctionnalités pourraient être ajoutées ou complétées:

- gestion avancée des droits utilisateurs avec par exemple la possibilité d'autoriser ou non la création de nouveaux comptes, ajouter une visibilité des collections différentes selon les rôles
- modification de la structure de layout: offrir la possibilité de modifier l'arbre formé par la structure du document
- annotation des entités nommées: permettre d'annoter un tag donné à plusieurs mots en même temps, permettre un retour en arrière en cas d'erreurs, ré-ordonnancement des labels pour un token donné

Libération du logiciel

Notre but est de libérer PIVAN et d'aider ses utilisateurs à se l'approprier. L'ingénieur devra donc développer une documentation utilisateur et surtout administrateur. Des partenaires historiens démographes ont manifesté leur intérêt pour PIVAN et l'ingénieur(e) sera leur interlocuteur privilégié pour le déploiement et l'utilisation de PIVAN sur leur infrastructure, ainsi que pour améliorer PIVAN et sa documentation en fonction de leurs retours. Nous ferons alors la publicité la plus large possible au logiciel.

Technologies utilisées: Java, React, Spring boot, SQL, IIIF
Compétences: Programmation Web, Programmation Objet, Communication, Rédaction
Durée: 12 mois
Lieu: Laboratoire LITIS, Université de Rouen Normandie, Saint-Étienne-du-Rouvray.
Contact : Thierry.Paquet@univ-rouen.fr Pierrick.Tranouez@univ-rouen.fr thomas.constum1@univ-rouen.fr