# PROPOSITION DE STAGE DANS LE CADRE DU PROJET AGODA

## 1/ PRÉSENTATION GÉNÉRALE DU STAGE

Le stage proposé s'inscrit dans le cadre du projet AGODA (*Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale)* qui associe le **Laboratoire de recherche historique Rhône-Alpes** (LARHRA) (CNRS-Lyon 2), le laboratoire **Méthodes Numériques en Sciences Humaines et Sociales** (MNSHS) à EPITA Paris, et **ALMAnaCH** (Automatic Language Modelling and Analysis & Computational Humanities) à Inria Paris.

AGODA est **l'un des cinq projets pilotes financés par le DataLab de la Bibliothèque nationale de France** pour l'année 2021-2023 (https://www.bnf.fr/fr/bnf-datalab). L'objectif du projet consiste à créer une **plateforme de consultation et d'exploration des débats parlementaires** retranscrits dans le *Journal officiel de la République française. Débats parlementaires. Chambre des députés : compte-rendu in-extenso* et disponibles sur Gallica (https://gallica.bnf.fr/ark:/12148/cb328020951/date.item). Cette plateforme de consultation sera basée sur l'application TEI Publisher (https://teipublisher.com/index.html), reposant sur une base de données XML (eXist-db). Dans cette première phase du projet, le ou la stagiaire sera amené.e à travailler sur une partie restreinte du corpus à savoir **les débats qui se sont tenus entre 1889 et 1893**.

Les débats parlementaires constituent une source incontournable pour l'histoire politique, mais aussi sociale, économique et culturelle, de la période contemporaine. Il s'agit également d'une source majeure pour l'histoire du droit. Pour ce faire AGODA entend développer un *workflow* permettant l'éditorialisation et l'enrichissement de gros corpus textuels historiques qui serait réutilisable pour d'autres projets.

Le stage aura pour objectif de contribuer à la **mise en place d'une chaîne de traitement des documents étudiés.** Il s'inscrira dans la continuité du travail réalisé par Fanny Lebreton lors de son stage de M2 effectué en 2022. Le mémoire et les livrables de ce stage sont disponibles en ligne sur Github : https://github.com/mpuren/agoda pour les livrables techniques et https://github.com/FannyLbr/Memoire-AGODA-TNAH2022 pour le mémoire.

Le ou la stagiaire participera à la préparation des textes en vue de leur publication des textes (**post-correction et annotation en TEI**) et à leur mise en ligne (**stockage des documents XML dans une base de données eXist-db**). Un autre versant du travail consistera à compléter la stratégie de **balisage automatique des fichiers textes** issus de l'OCRisation, à partir d'un schéma XML adapté aux débats parlementaires et formalisé par un fichier ODD documenté. Pour ce faire, il conviendra d'ajuster et de compléter un ensemble de scripts Python. Il conviendra également de produire une documentation décrivant la stratégie et les méthodes employées et pouvant être utilisée par de futurs utilisateurs. Enfin, le ou la stagiaire pourra contribuer à la mise en ligne des textes annotés en les stockant dans une base de données eXist-db.

Le stage nécessitera donc une bonne connaissance du **XML-TEI**, de solides compétences en programmation **Python** et, éventuellement, en XSLT ainsi qu'une capacité à travailler en équipe et à organiser rigoureusement son travail.

Le ou la stagiaire aura pour mission de :

- Modifier et compléter la documentation du schéma XML développé (comportement d'éléments à modifier, ajout de nouveaux éléments TEI, etc.) et formalisé par un fichier ODD documenté ;
- Mettre en place la vérification automatique de la conformité des fichiers TEI produits au schéma utilisé, par exemple avec la librairie Python lxml ;
- Compléter l'automatisation du processus de balisage : faire monter le processus en généricité pour l'appliquer à un corpus plus large ;
- Contribuer à la création de la vérité de terrain en utilisant l'outil d'OCR BibliLense développé par EPITA. Les données seront mises à disposition de façon ouverte (FAIR) avec un dépôt sur HTR-United (https://htr-united.github.io/) ;
- Contribuer à l'amélioration du processus de post-correction des fichiers issus de l'OCR ;
- Contribuer au développement de BibliLense en vue de la production de données en XML-TEI à partir des documents océrisés ;
- Contribuer au développement de l'annotation sémantique des documents via BibliLense notamment des entités nommées avec un modèle de langue type BERT ;

## 2/ Condition du stage

Le stage pourra se dérouler soit dans les **locaux du laboratoire LRE (14-16 rue Voltaire au Kremlin-Bicêtre)**, soit dans les l**ocaux dans les locaux du LARHRA, 14 avenue Berthelot à Lyon** (au choix du ou de la stagiaire). S'il ou elle le souhaite, le ou la stagiaire pourra également travailler dans les espaces du DataLab qui sont réservés aux participants aux projets financés (BnF, site Tolbiac – François Mitterrand).

Il est également possible d'envisager qu'une partie du stage se déroule à distance (au choix du ou de la stagiaire).

Le stage doit débuter en avril 2023, pour une durée de 4 mois. Le ou la stagiaire bénéficiera de la gratification de stage minimale prévue par la réglementation : https://www.service-public.fr/particuliers/vosdroits/R40280

## 3/ Personnes ressources

Le ou la stagiaire bénéficiera du suivi des chercheurs.ses et ingénieurs.res du LARHRA, d'EPITA et d'Inria. Il ou elle sera également suivi.e par le DataLab, ainsi que par Huma-Num, l'infrastructure de recherche dédiée aux Humanités numériques.

Technique

A EPITA : **Aurélien Pellet**, ingénieur de recherche (océrisation des textes et de post-correction). Mail : : aurelien.pellet @epitech .eu

**Edwin Carlinet**, enseignant-chercheur (outil BibliLense). Mail : edwin1.carlinet@epita.fr

**Joseph Chazalon**, enseignant-chercheur (outil BibliLense). Mail : josep.chazalon@epita.fr

Au LARHRA : **Morgane Pica**, ingénieure d'études (annotation des textes en TEI). Mail : morgane.pica@ens-lyon.fr

Au DataLab Bnf et à Huma-Num : **Margaux Faure**, ingénieur d'études (chargé et du suivi technique de toute la chaîne de traitement). Mail : margaux.faure@huma-num.fr

A INRIA :

**Eric de la Clergerie**, chargé de recherche en TAL (annotation des textes en TEI). Mail : Eric.de_la_clergerie@inria.fr

**Julien Martin**, développeur (développement de la plateforme TEI Publisher). Mail : julien.pierre.martin@gmail.com

Scientifique

À EPITA: **Marie Puren**, enseignante-chercheuse à EPITA en histoire et humanités numériques (coordinatrice du projet). Mail : marie.puren@epita.fr

Au LARHRA : **Pierre Vernus**, MCF à l'Université Lumière Lyon 2 en histoire contemporaine (coordinateur du projet). Mail : pierre.vernus@msh-lse.fr

## 3/ Autres informations sur le stage

S'il ou elle le souhaite, le ou la stagiaire pourra être associé.e en tant qu'auteur.e à la rédaction des publications scientifiques issus du projet (communications dans des colloques et articles scientifiques). Ces publications seront évidemment rédigées en étroite collaboration avec les autres participant.e.s du projet.

Par ailleurs, s'il ou elle le souhaite, le ou la stagiaire pourra présenter les résultats de ses travaux durant des colloques et conférences,

Pour en savoir plus sur le projet, on peut consulter : https://hal.science/hal-03382765

Pour un exemple de travail scientifique réalisé sur ces documents, on pourra également consulter : https://hal.archives-ouvertes.fr/hal-03526254.