# Génération de manuels scolaires à partir de données annotées

## **Contexte**

Le projet ANR MALIN a pour objectif de rendre utilisables les manuels scolaires numériques par les enfants en situation de handicap dont l’inclusion dans les établissements scolaires ordinaires a été posée par la loi du 11 février 2005. En France, le manuel scolaire est un support pédagogique emblématique quasi systématiquement utilisé en classe afin d’accompagner de manière progressive et structurée l’acquisition des connaissances et compétences définies dans les programmes scolaires officiels de l’Éducation Nationale. Les manuels numériques actuellement disponibles nécessitent d’être adaptés pour être utilisés par les élèves en situation de handicap. Ces adaptations concernent aussi bien les aspects techniques que pédagogiques. Dans la plupart des cas, les manuels sont adaptés de façon artisanale et les délais de livraison peuvent être de plusieurs mois. Ces contraintes ne permettent pas de rendre efficiente l’inclusion scolaire. L’objectif du projet ANR MALIN est donc de développer des solutions techniques afin d’aboutir à **l’automatisation de l’adaptation des manuels scolaires numériques pour les rendre accessibles** (accès, traitement et interaction avec les contenus) **aux élèves en situation de handicap**.

Ce stage s’inscrit dans le cadre d’un projet ANR.

Le projet ANR repose sur une collaboration entre quatre laboratoires : LISN (Université Paris Saclay), MISC (Ecole CentraleSupelec), CEDRIC (CNAM), Inserm 1284 (CRI, Université de Paris). Le stagiaire travaillera en interaction avec deux doctorants associés au projet.

## **Objectif du stage**

L’objectif général de ce stage est la génération automatique de manuels scolaires ou de parties de manuels scolaires pour pouvoir entraîner des modèles d’apprentissage profond à analyser, comprendre et adapter ces manuels. Il s’agit d’une tâche complexe car en plus de la nature multimodale des éléments d’un manuel (texte, bloc de textes, exercices, leçons, consignes, image, graphique, tableau ...) il est aussi nécessaire de comprendre la structure et la mise en page du document.

Des avancées considérables ont été faites sur cette problématique (Visual Document Understanding), notamment à l’aide d’architectures de type transformers comme DocFormer **[Appalaraju et al, 2021]** ou XYLayoutLM **[Gu et al, 2022].** Ces modèles sont pré-entrainés de manière auto-supervisée et peuvent être utilisés par fine-tuning pour d’autres tâches, notamment de compréhension du document. Cette dernière étape nécessite notamment d’avoir des jeux de données labelisées en suffisamment grande quantité, ce qui n’est pas réalisable dans le cadre du projet MALIN. Pour répondre à ce manque de données, une des approches prometteuses de la littérature est la génération de données **[Besnier et al, 2020].** Les modèles génératifs, comme les GANs ou les modèles de diffusion **[Yang et al, 2022]** pour les images ou les modèles de type GPT pour le texte sont en effet capables de générer des données très réalistes aujourd’hui.

Dans ce contexte, il est attendu du stage d’une part de faire un état des lieux des modèles de génération de documents multimédia,de choisir et de mettre en œuvre un ou plusieurs modèles pour la génération de manuels scolaires.

**[Appalaraju et al, 2021]** Appalaraju, Srikar and Jasani, Bhavan and Kota, Bhargava Urala and Xie, Yusheng and Manmatha, R. ; DocFormer: End-to-End Transformer for Document Understanding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021

**[Gu et al, 2022]** Gu, Zhangxuan and Meng, Changhua and Wang, Ke and Lan, Jun and Wang, Weiqiang and Gu, Ming and Zhang, Liqing; XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding; Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2022

[**Besnier et al, 2020]** Besnier, Victor & Jain, Himalaya & Bursuc, Andrei & Cord, Matthieu & Perez, Patrick. (2020). This Dataset Does Not Exist: Training Models from Generated Images – ICASSP 2020

**[Yang et al, 2022]** Ling Yang and Zhilong Zhang and Shenda Hong and Runsheng Xu and Yue Zhao and Yingxia Shao and Wentao Zhang and Ming-Hsuan Yang and Bin Cui; Diffusion Models: A Comprehensive Survey of Methods and Applications; ArXiv,

## **Compétences**

* master et/ou ingénieur en informatique avec une spécialisation en apprentissage automatique et traitement automatique de la langue
* maîtrise de Python (langage du projet)
* excellent niveau de langue française

## **Informations générales**

Lieu de travail : Paris

Durée du contrat : 4 à 6 mois

Date d'embauche souhaitée : à partir de avri/mail 2023

Quotité de travail : Temps complet

Gratification de stage organisme public

Niveau d'études souhaité : Bac+5

## **Contact :** [**olivier.pons@lecnam.net**](mailto:olivier.pons@lecnam.net)[**vincent.mousseau@centralesupelec.fr**](mailto:vincent.mousseau@centralesupelec.fr) **caroline.huron.unicog@gmail.com**