# Extraction de contenu multimodal
# Application au cas des manuels scolaires

## Contexte

Le projet ANR MALIN a pour objectif de rendre utilisables les manuels scolaires numériques par les enfants en situation de handicap. En effet, les manuels numériques actuellement disponibles nécessitent d'être adaptés pour être accessibles à ces enfants. 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 des enfants en situation de handicap. 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**. Le projet ANR repose sur une collaboration entre quatre laboratoires : LISN (Université Paris-Saclay), MICS (Ecole CentraleSupelec), CEDRIC (CNAM), Inserm 1284 (CRI, Université de Paris).

## Sujet du stage

L'objectif du stage consiste à concevoir des approches d'extraction automatique de la structure d'un exercice de manuel scolaire (consignes, enoncés, exemples, etc.) et de son contenu multimédia (textes, images, dessins, graphiques, équations, courbes...) à partir des fichiers fournis par les éditeurs (ceux-ci sont le plus souvent au format pdf). Plusieurs approches seront à envisager : une approche d'adaptation et d'enrichissement de systèmes de structuration automatique de documents textuels (segmentation thématique, segmentation discursive) prenant en compte la spécificité et la multi-modalité des données traitées et une approche basée sur le traitement automatique des images visant à identifier les différents blocs en se basant sur les caractéristiques de l'image, connue sous le nom de « Document Layout Segmentation and Analysis » [1, 2]. Des approches récentes d'apprentissage profond seront testées sur des jeux de données annotées manuellement afin d'adapter des modèles existants et obtenir des résultats d'extraction satisfaisants.

## Compétences

- Master en informatique ou TAL avec une spécialisation dans au moins un des domaines suivants :
    - traitement automatique des langues
    - apprentissage automatique
- Maîtrise de Python (langage de prédilection du projet)

La connaissance des principales librairies d'apprentissage sera appréciée.

## Informations générales

Lieu de travail : Laboratoire CEDRIC du CNAM
Durée du contrat : 5/6 mois
Début souhaité : printemps 2024
**Contact :** Pour postuler, merci d'envoyer un CV, les notes de M1 et M2 et une lettre de motivation à Camille Guinaudeau (guinaudeau@limsi.fr), Olivier Pons (olivier.pons@lecnam.net) et Caroline Huron (caroline.huron@cri-paris.org).

## Références

[1] Wang, Jiapeng, Lianwen Jin, and Kai Ding. "Lilt: A simple yet effective language-independent layout transformer for structured document understanding." arXiv preprint arXiv:2202.13669 (2022).
[2] Huang, Yupan, et al. "Layoutlmv3: Pre-training for document ai with unified text and image masking." Proceedings of the 30th ACM International Conference on Multimedia. 2022.