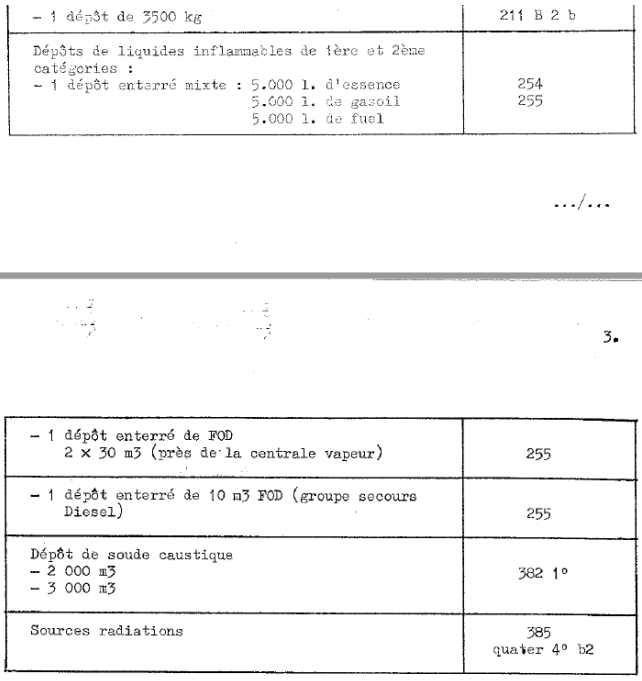
Projet 1 – Détection de blocs tableaux

L'objectif de ce projet est de construire un programme qui permette de découper des arrêtés préfectoraux scannés pour en séparer les *blocs textes* et les *blocs tableaux*.

Les tableaux peuvent se présenter dans différentes conditions :

- Apparaître en entier sur une page
- Apparaître coupés entre plusieurs pages
- Être précédés ou suivis de texte

Exemple d'entrée :



– 1 dépôt de 3500 kg	211 B 2 b
Dépôts de liquides inflammables de 1ère et 2ème catégories : – 1 dépôt enterré mixte : 5.000 l. d'essence 5.000 l. de gasoil 5.000 l. de fuel	 254 255

.../...

3.

– 1 dépôt enterré de FOD 2 × 30 m3 (près de la centrale vapeur)	255
– 1 dépôt enterré de 10 m3 FOD (groupe secours Diesel)	255
Dépôt de soude caustique – 2 000 m3 – 3 000 m3	382 1°
Sources radiations	385 quater 4° b2

Pour les tableaux coupés entre plusieurs pages, il conviendra de détecter que les différents blocs tableau ne sont en réalité pas plusieurs tableaux distincts, mais bien plusieurs parties du même tableau.

Pour chaque bloc, il conviendra, en plus de son type, de conserver aussi les métadonnées suivantes :

- Le numéro de page dans le document d'origine
- Les coordonnées de la bounding box du bloc dans cette page
- L'index global de ce bloc dans le document d'origine (dans le sens de lecture)

Exemple de sortie :

```
blocs = [
  ‘’tableau’’ : {
    ‘’page_number’’ : 2,
    ‘’coords_start’’ : 210.7,
    ‘’coords_end’’ : 388.2,
  },
```

```
  ''tableau'' : {
    ''page_number'' : 3,
    ''coords_start'' : 0,
    ''coords_end'' : 110.3,
  },
]
```

Les blocs extraits par votre programme pourront être sauvés en format image. Vous devrez vous assurer que les tableaux seront entièrement visibles bordures incluses dans les limites définies par vos bounding boxes, sans mordre sur d'autres blocs. Les métadonnées associées à ces images seront sauvegardées dans un format libre.

Bonus

Une fois les blocs tableau sauvés au format image, vous pouvez essayer différentes techniques de reconnaissance optique de caractères (OCR). Votre solution devra présenter le maximum d'informations liées au tableau : mise en forme, cellules fusionnées, texte complet et écrit dans les bonnes cases.