Séance 8

Corpus TEI

Traitement automatisé d’un corpus en TEI

Pratiques de la TEI

Encodage manuel / encodage automatique

Envergure de l’encodage

Texte unique

Encodage fait dans le détail nécessitant une expertise sur le texte et contexte

Analyse fine du contenu

Corpus élargi

Encodage qui signale la structuration générale et les éléments principaux du contenu

Aperçu global du corpus

Automatisations de l’encodage

Balisage avec ReGex

Ajout de balises à l’aide du formatage initial du texte à encoder

Transformation XSLT

Transformation d’un fichier XML en un fichier TEI enrichi

Langage de script

Utilisation de script (e.g. Python) pour composer un fichier TEI à partir d’une ou plusieurs sources

Exercice

Ajout de balises avec ReGeX

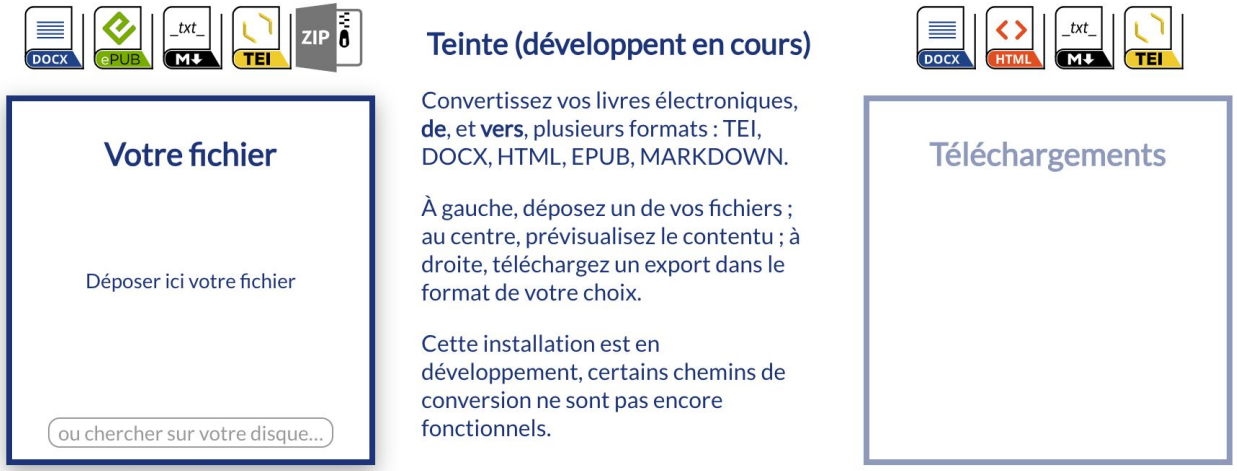
Consigne

À partir de la transcription du Cid :

1. Baliser les actes :
   `^\t{4}(ACTE (.[^\n]*))$`
2. Baliser les scènes :
   `^\t{4}(SCÈNE (.[^\n]*))\.$`
3. Baliser les prises de paroles :
   `^\t{4}([A-Z’]{2,}[^\n]*)\.\n\t{4}(.+?)\n\t{4}\n\t{4}`

Transformation fichier binaire vers TEI avec TEINTE

obtic.huma-num.fr/teinte



Outils pour la publication et le traitement TEI

TEI Publisher

Plateforme de mise en ligne et présentation de corpus TEI

Obtic (Sorbonne Univ.)

Série d'outils développés dans le laboratoire pour le traitement des textes et corpus

eScriptorium (EPHE)

Interface de transcription de document manuscrits assistée par HTR et reconnaissance de structure

Pyrrha (ENC)

Interface de correction de corpus lemmatisés et annotés automatiquement d'un point de vue morpho-syntaxique

Bonnes pratiques

Dans la phase de modélisation, rechercher des technologies/stratégies éprouvées :

ODD déjà formalisées, chaînes de traitement pour textes/problématiques similaires, standards de métadonnées, ontologies, etc.

Conception d’une chaîne de traitement

- Garder les **fichiers sources** et une trace des différents traitements
- Prévoir un niveau d’**encodage minimal** à atteindre pour tous les documents
- `encodingDesc` et `revisionDesc` peuvent être utilisés pour détailler les **étapes de traitement** de l’édition/corpus

Exercice

Interopérabilité des métadonnées

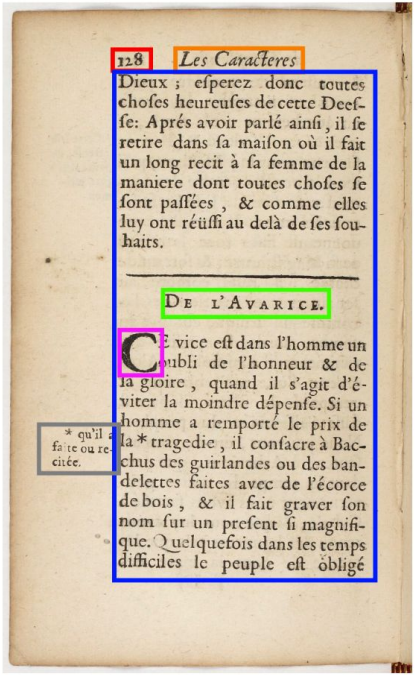
Dublin Core

Consigne

À partir de la transcription du Cid, remplir le TEI header avec les métadonnées Dublin Core :

Title	Creator	Subject
Description	Publisher	Contributor
Date	Type	Format
Identifier	Source	Language
Relation	Coverage	Rights

Segmentation automatique & HTR



Page XML / ALTO

segmonto.github.io

Objectif de l’automatisation

Avec des corpus numérisés de plus en plus important et l’accès à des outils d’extraction de texte de plus en plus perfectionnés (OCR / HTR), l’enjeu consiste à **enrichir ces ressources**

Organisation du corpus

Fichier unique

Fichier TEI dont le `<body>` se compose d'un ensemble de sous-unités structurées selon un même modèle

Adapté à des corpus homogène de textes présentant des métadonnées descriptives communes

Dossiers

Ensemble de fichiers indépendants, structurés dans une arborescence contrôlée avec convention de nommage

Utile dans des chaînes de traitement avec variété de formats de fichiers et/ou grande diversité de textes

`teiCorpus`

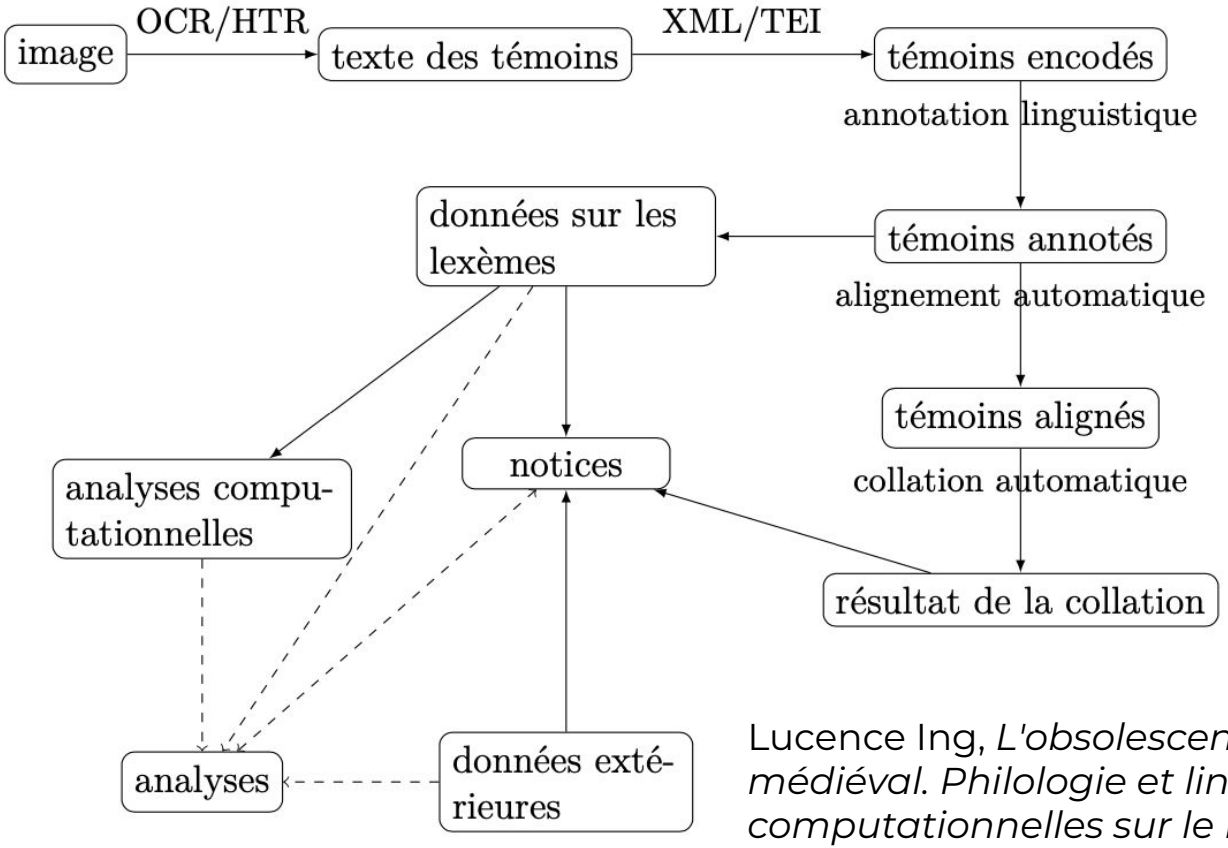
Fichier TEI composé d'une imbrication de documents TEI possédant chacun leur `<teiHeader>`

Utile lorsque les textes encodés possèdent des métadonnées individuelles malgré la cohérence de l'ensemble

Fichier unique

Corpus homogène

Chaîne de traitement pour fiches lexicales



Lucence Ing, *L'obsolescence lexicale en français médiéval. Philologie et linguistique computationnelles sur le Lancelot en prose*, 2023

teiCorpus

Corpus de textes composites

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title></title>
         </titleStmt>
         <publicationStmt>
            <p></p>
         </publicationStmt>
         <sourceDesc>
            <p></p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <p></p>
      </body>
   </text>
</TEI>
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title></title>
         </titleStmt>
         <publicationStmt>...</publicationStmt>
         <sourceDesc>...</sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <group>
         <text><body><p></p></body></text>
         <text><body><p></p></body></text>
         <text><body><p></p></body></text>
         ...
      </group>
   </text>
</TEI>
```

```
<teiCorpus version="3.3.0" xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <!--[en-tête du corpus]-->
    </teiHeader>
    <TEI>
        <teiHeader>
            <!--[en-tête du premier texte]-->
        </teiHeader>
        <text>
            <!--[premier texte du corpus]-->
        </text>
    </TEI>
    <TEI>
        <teiHeader></teiHeader>
        <text></text>
    </TEI>
</teiCorpus>
```

Répertoire Sévigné

Bienvenue sur le site consacré à la correspondance de Madame de Sévigné, créé et animé par Cécile Lignereux, Simon Gabay et ELAN, l'équipe Littératures et Arts Numériques de l'UMR 5316 (Litt&Arts).

Vous y trouverez un répertoire des sources imprimées et manuscrites, une liste des travaux critiques régulièrement mise à jour, ainsi que des actualités et des ressources iconographiques.

REPERTOIRE. s. m. Inventaire, table ou recueil, où les choses, les matieres sont rangées dans un ordre, qui fait qu'on les trouve facilement. (Furetière, *Dictionnaire universel*, 1727).

gitlab.com/litt-arts-num/madame-de-sevigne

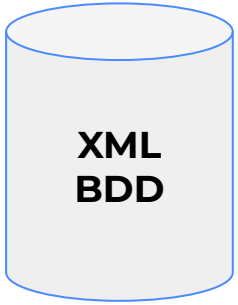
```
<teiCorpus version="3.3.0" xmlns="http://www.tei-c.org/ns/1.0"
           xmlns:xi="http://www.w3.org/2001/XInclude">
   <teiHeader>...</teiHeader>
   <xi:include href="HN_Sevigne_index.xml"/>
</teiCorpus>

HN_Sevigne_index.xml
<TEI xml:id="index" xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>...</teiHeader>
   <text></text>
</TEI>
```

Fichiers structurés

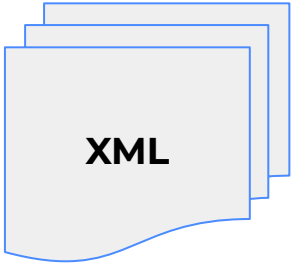
DIfférents fichiers TEI organisés selon une même chaîne de traitement

Stockage des fichiers



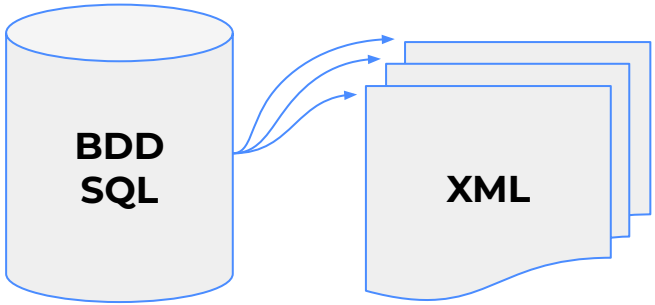
Base de données XML

BaseX, eXistDB



Arborescence de fichiers

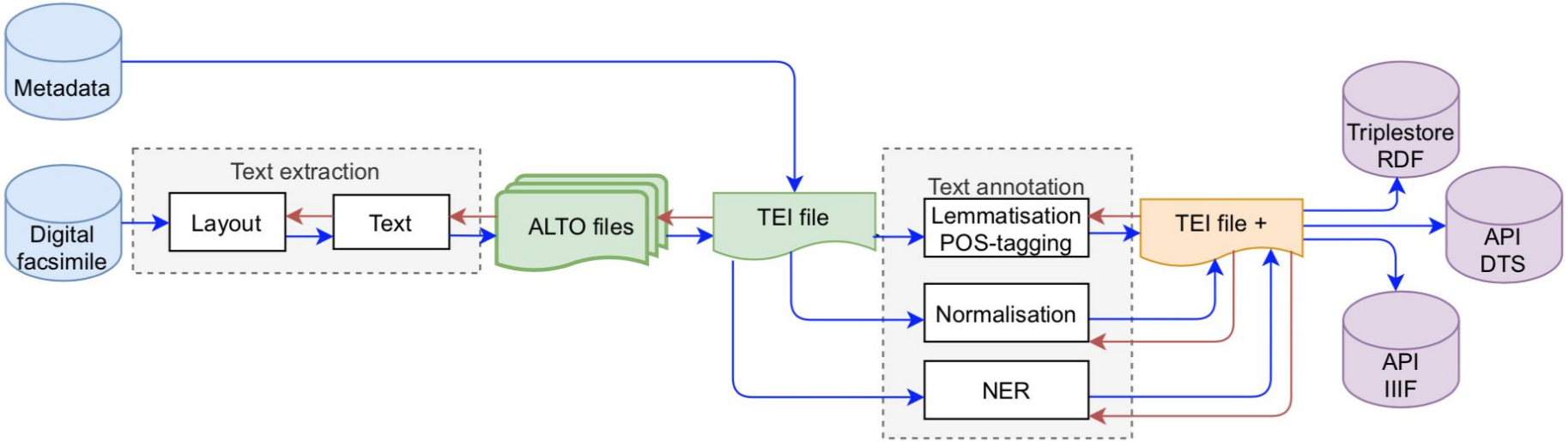
Organisation en dossiers + convention de nommage



Stockage mixte

Base relationnelle classique dont une ou plusieurs tables présentent des champs faisant référence à un chemin de fichier XML

Chaîne de traitement pour Galli(corpor)a



Kelly Christensen, *D'ALTO à TEI, modélisation de transcriptions automatiques pour une pré-éditorialisation des textes*, 2022

Quelques exemples de projets

Gallic(orpor)a

Corpus en diachronie longue: *training data* enrichissement et valorisation des ressources Gallica

COREL

Recomposition de la législation chinoise à partir d'un corpus morcelé et hétérogène

AGODA

Mise à disposition des transcriptions de débats parlementaires de la chambre des députés, 1881-1940

LECTAUREP

Repenser la mise à disposition et exploitation des archives notariales

Time-US

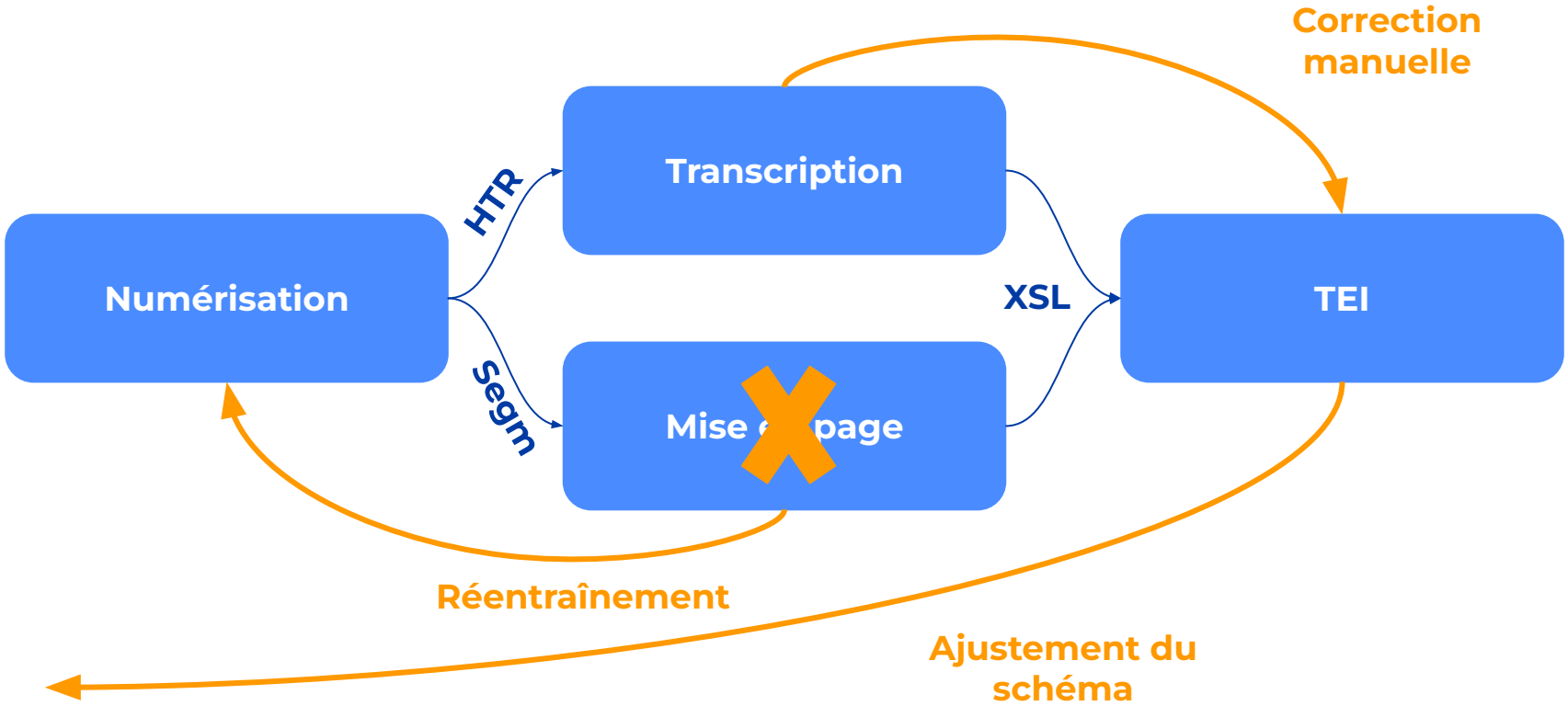
Reconstituer les rémunérations et temps de travail des travailleur·se·s et des travailleurs du textile

CoMeRe

Corpus de genres de communication médiatisée par ordinateur (CMC) d'interactions en langue française

Ça semble simple

Beaucoup d’essai/erreur



Segmentation et premiers échecs

Le modèle de segmentation *Kraken* ne nous a pas permis d'obtenir des résultats satisfaisants. Problème d'hétérogénéité des mises en page ? De granularité de la description ? Ou problème méthodologique ?

FIGURE – BnF, Réserve des livres rares, vélins 611, 15e s.

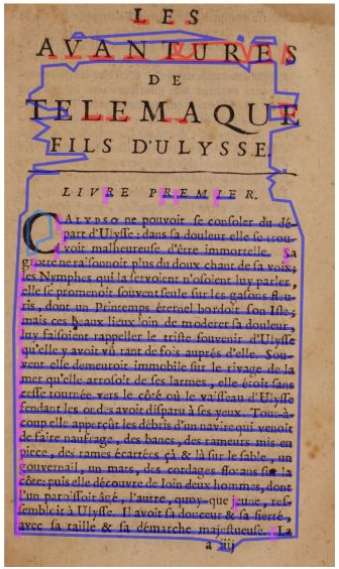


FIGURE – BnF, Réserve des livres rares, RES-Z-2442, 16e s.

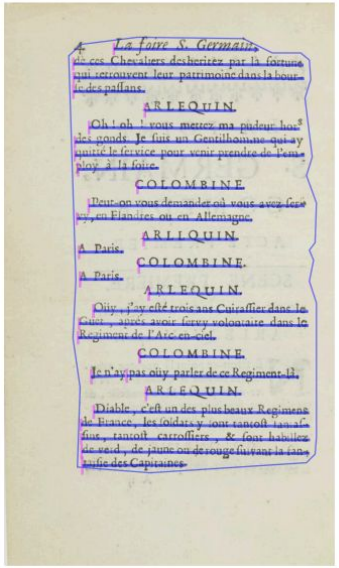


FIGURE – BnF, Arts du spectacle, Réserve 8-RO-1702, 17e s.

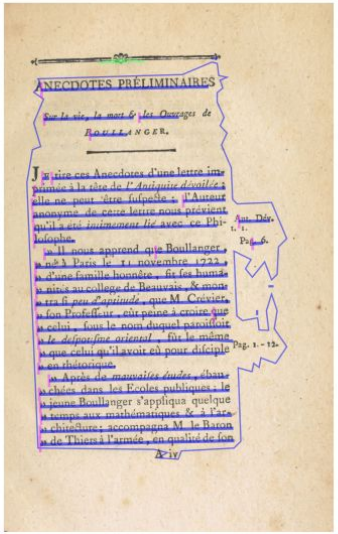


FIGURE – BnF, département Droit, économie, politique, 2012-39571, 18e s.

Conception du modèle textuel

Similitudes & divergences

Identifier les informations à encoder et voir où celles-ci sont situées dans les textes

Balisage d'une version

Choix des éléments et de la structure pour l'encodage d'une version

Génération des specifications

Création des règles des validations qui serviront de guide pour l'ensemble des encodages

Exercice

Repérer les variations des textes à encoder

Consigne

À partir des transcriptions de lettres :

1. Identifier les informations importantes à encoder
2. Relever les différences entre les différentes lettres
3. Créer un modèle d’encodage pour l’ensemble des lettres
4. Encoder les autres lettres selon le même modèle