

# Introduction à la TEI : Créer son édition scientifique numérique

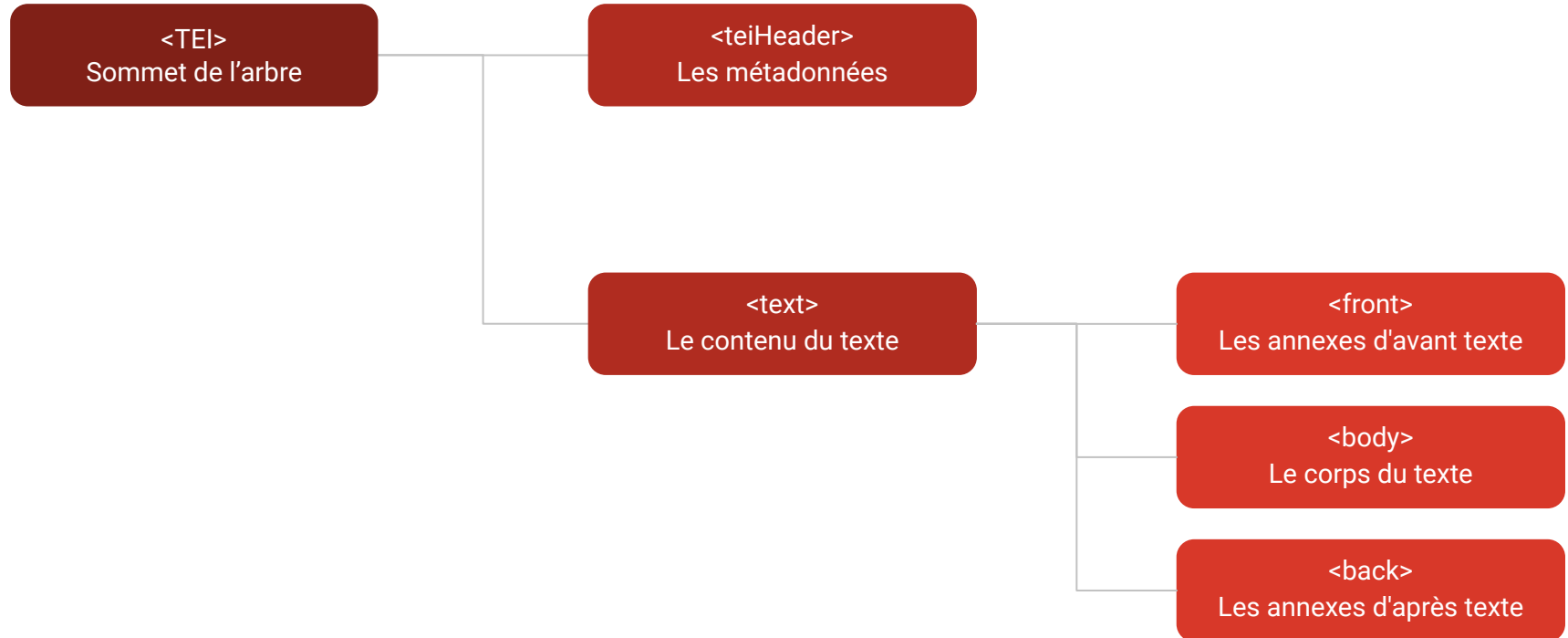
Floriane Chiffolleau, Doctorante en Humanités Numériques, ALMAAnaCH/Le Mans Université

URFIST de Rennes, 24-25 novembre 2022

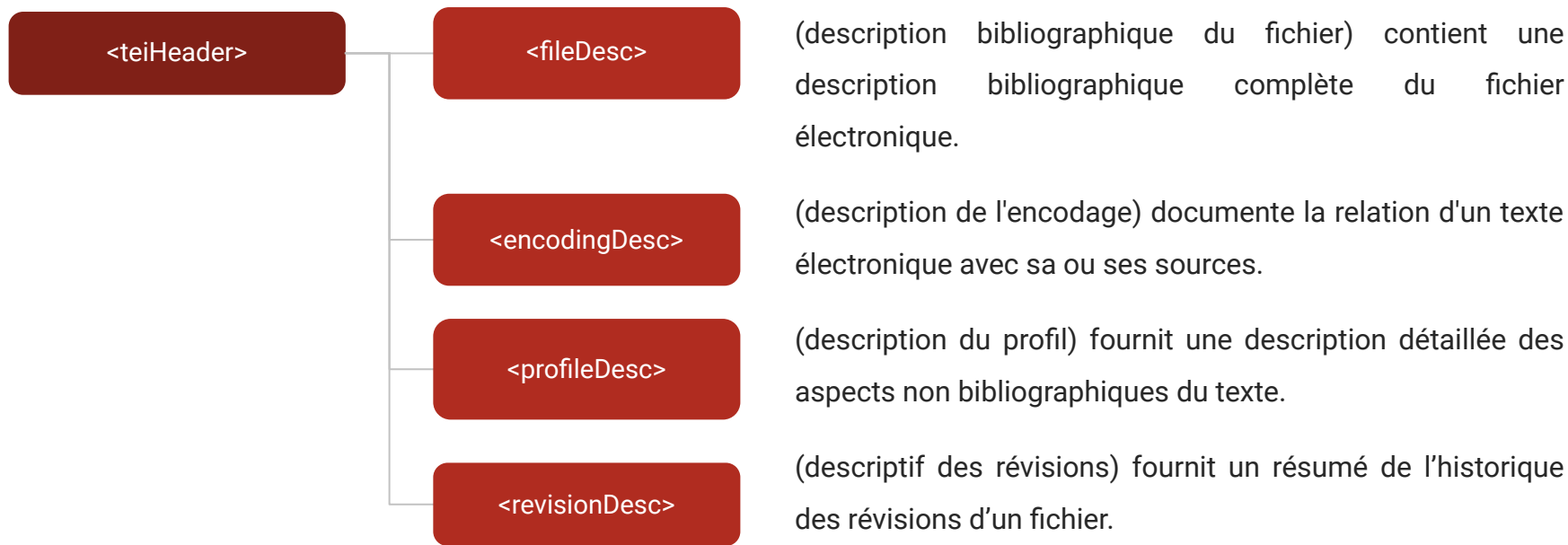
Repository GitHub : [https://github.com/FloChiff/Introduction\\_TEI\\_2022](https://github.com/FloChiff/Introduction_TEI_2022)

# La TEI au service de son édition

# Contenu de l'arbre XML-TEI



# Les métadonnées : le <teiHeader>



# Les métadonnées : le <teiHeader>

## Chapitre 2 The TEI Header

<https://tei-c.org/release/doc/tei-p5-doc/fr/html/HD.html>

### Le <fileDesc>

- ❑ **<titleStmt>** (mention de titre) regroupe les informations sur le titre d'une œuvre et les personnes ou institutions responsables de son contenu intellectuel.
- ❑ **<editionStmt>** (mention d'édition) regroupe les informations relatives à l'édition d'un texte.
- ❑ **<publicationStmt>** (mention de publication) regroupe des informations concernant la publication ou la diffusion d'un texte électronique ou d'un autre type de texte.

# Les métadonnées : le <teiHeader>

## Le <fileDesc>

- ❑ **<seriesStmt>** (mention de collection) regroupe toute information relative à la collection (si elle existe) à laquelle appartient une publication.
- ❑ **<notesStmt>** (mention de notes) rassemble toutes les notes fournissant des informations sur un texte, en plus des informations mentionnées dans d'autres parties de la description bibliographique.
- ❑ **<sourceDesc>** (description de la source) décrit la source à partir de laquelle un texte électronique a été dérivé ou produit, habituellement une description bibliographique pour un texte numérisé ou une expression comme "document numérique natif " pour un texte qui n'a aucune existence précédente.

# Les métadonnées : le <teiHeader>

## L'<encodingDesc>

- ❑ Décrire l'encodage du fichier électronique : description du projet <projectDesc>, déclaration d'échantillonnage <samplingDecl>, déclaration des pratiques éditoriales <editorialDecl> et autres types de déclarations (balisage, classifications, coordonnées géographiques, etc).

## Le <profileDesc>

- ❑ Fournir la description détaillée des aspects non bibliographiques : date de création du texte <creation>, langues utilisées dans le texte <langUsage>, etc.

## Le <revisionDesc>

- ❑ Fournir un résumé de l'historique de révisions d'un fichier. Encodage avec la balise <change> et à l'intérieur les attributs @who (responsable de la modification) et @when (date de la modification).
- ❑ Encodage à faire dans un ordre chronologique inversé (plus récent en premier)

# Les métadonnées : le <teiHeader>

Le <fileDesc>

## ❏ <titleStmt>

```
<titleStmt>
  <title type="main">Orgueil et préjugé</title>
  <title type="sub">Par l'auteur de Raison et Sensibilité</title>
  <title type="sub">Traduit de l'anglais</title>
  <author>Jane Austen</author>
  <respStmt>
    <resp>Transcription by</resp>
    <name>
      <forename>Floriane</forename>
      <surname>Chiffoleau</surname>
    </name>
  </respStmt>
  <respStmt>
    <resp>Encoded by</resp>
    <name>
      <forename>Floriane</forename>
      <surname>Chiffoleau</surname>
    </name>
  </respStmt>
</titleStmt>
```



# Les métadonnées : le <teiHeader>

## Le <fileDesc>

### ❑ <publicationStmt>

```
<publicationStmt>
  <publisher>J. J. Paschoud, libraire</publisher>
  <pubPlace>Paris</pubPlace>
  <date>1822</date>
</publicationStmt>
```

### ❑ <seriesStmt>

```
<seriesStmt>
  <title type="main">Littérature anglaise</title>
</seriesStmt>
```

### ❑ <sourceDesc> → nativement numérique

```
<sourceDesc>
  <p>Born digital</p>
</sourceDesc>
```

### ❑ <sourceDesc> → extrait d'un manuscrit

```
<sourceDesc>
  <msDesc>
    <msIdentifier>
      <country>France</country>
      <settlement>Paris</settlement>
      <institution>Bibliothèque nationale de France</institution>
      <collection>Réserve des livres rares</collection>
      <idno>RES P-YZ-2802</idno>
    </msIdentifier>
  </msDesc>
</sourceDesc>
```

# Les métadonnées : le <teiHeader>

## L'<encodingDesc>

### ❑ <projectDesc>

```
<projectDesc>
  <p>L'encodage de ce document s'est fait dans le cadre d'une formation d'introduction à la TEI.</p>
</projectDesc>
```

### ❑ <samplingDecl>

```
<samplingDecl>
  <p>Seuls les textes de littérature anglaise du début du XIXème siècle ont été choisis.</p>
</samplingDecl>
```

### ❑ <editorialDesc>

```
<editorialDecl>
  <correction>
    <p>Aucune correction</p>
  </correction>
  <hyphenation eol="hard" rend="sh">
    <p>Toutes les coupures de mots pour fin de ligne (indiquées principalement avec un tiret simple) ont été conservées.</p>
  </hyphenation>
</editorialDecl>
```

# Les métadonnées : le <teiHeader>

Le <profileDesc>

❏ <particDesc>/<settingDesc>

```
<particDesc>
  <listPerson>
    <person xml:id="p0001">
      <persName>
        <forename>Elisabeth</forename>
        <surname>Bennet</surname>
      </persName>
    </person>
    <person xml:id="p0002"> [5 lines]
    <person xml:id="p0003"> [5 lines]
    <person xml:id="p0004"> [2 lines]
  </listPerson>
</particDesc>
```

```
<settingDesc>
  <listPlace>
    <place xml:id="l0001" type="house">
      <placeName>Pemberley</placeName>
      <region>Derbyshire</region>
      <country>England</country>
    </place>
    <place xml:id="l0002" type="city"> [4 lines]
    <place xml:id="l0003" type="inn"> [5 lines]
  </listPlace>
</settingDesc>
```

# Les métadonnées : le <teiHeader>

Le <profileDesc>

❏ <creation>/<langUsage>

```
<creation when="1813"/>
<langUsage>
  <language ident="fr"/>
  <!--
    <language ident="fr" usage="75">Français</language>
    <language ident="en" usage="25">Anglais</language>
  -->
</langUsage>
```

Le <revisionDesc>

```
<revisionDesc>
  <change when-iso="2021-02-17" who="#floriane.chiffolleau">Added a particDesc and settingDesc in the profileDesc</change>
  <change when-iso="2021-02-16" who="#floriane.chiffolleau">Encoding of the file</change>
</revisionDesc>
```

# Les annexes : les <front> et <back>

Chapitre 4 Default Text Structure <https://tei-c.org/release/doc/tei-p5-doc/fr/html/DS.html>

## Éléments contenus dans le <front>

- ❑ Une **préface** ou un avant-propos adressé aux lecteurs par l'auteur
- ❑ Des **remerciements** rédigés par l'auteur
- ❑ Une ou des **dédicace(s)** faites par l'auteur pour des personnes en particulier ou une/des institution(s)
- ❑ Un **résumé** ou court extrait du contenu du texte
- ❑ Une **table des matières**, qui présente le contenu et la manière dont il est structuré au sein du texte
- ❑ Un **frontispice** avec la description, s'il y en a une, de l'illustration présente avec la page de titre

## Éléments contenus dans le <back>

- ❑ Une **annexe** contenant des sections auxiliaires du texte avec des informations complémentaires
- ❑ Un **glossaire** avec une liste de termes associée à leurs définitions
- ❑ Un regroupement de **notes** textuelles ou autre genre de notes contenues dans le texte
- ❑ Une liste des **citations bibliographiques** qui se retrouvent dans le texte
- ❑ Toute forme d'**index** associé à l'œuvre
- ❑ Un **colophon**, soit une note finale précisant les conditions physiques de production de l'ouvrage

# Le corps du texte : le <body>

Chapitre 3 Elements Available in All TEI Docs <https://tei-c.org/release/doc/tei-p5-doc/fr/html/CO.html#COPU>

Chapitre 4 Default Text Structure <https://tei-c.org/release/doc/tei-p5-doc/fr/html/DS.html>

Encodage du <body> à l'aide de divisions et subdivisions en utilisant la balise <div>. Trois attributs peuvent principalement être utilisés avec :

- ❑ @xml:id → donner un attribut bien spécifique à la division pouvant être référencé par la suite.
- ❑ @type → spécifier le type de divisions que l'on a (par exemple, tome, chapitre, poème, lettre, etc.)
- ❑ @n → numérotation et associé généralement au type que l'on aura défini avant

Possible de mettre un en-tête en début de division (titre de section, intitulé de liste, description de manuscrit, etc.) avec la balise <head>.

Deux autres balises importantes à retenir : saut de ligne <lb/> et saut de page <pb/> balises autofermantes

# Le corps du texte : le <body>

Diverses balises en fonction du type de texte encodé :

- ❑ **<p>** → marquer les paragraphes dans des textes en prose
- ❑ **<lg>** (groupe de vers) et **<l>** (vers) → encoder les poèmes ou les pièces de théâtre
- ❑ **<list>** (liste) et **<item>** (composant) → encoder une liste et ses composants
- ❑ **<table>** (tableau), **<row>** (ligne) et **<cell>** (cellule) → encoder le contenu d'un tableau
- ❑ **<figure>** (figure) → encoder une illustration. Elle s'accompagne généralement d'une balise **<graphic>** qui contiendra le lien URL vers l'image. Elle peut aussi être accompagnée d'un **<figureDesc>** qui servira à décrire l'image.

Bien d'autres balises existent pour encoder les textes, de manière plus ou moins détaillée, en fonction du type de texte sur lequel on travaille (une pièce de théâtre, un roman, un poème, un discours, etc.). Quelques chapitres des guidelines sont dédiés exclusivement à certains de ces textes.

# Le corps du texte : le <body>

## Exemples

Un extrait de roman : <div> imbriqués contenant des attributs et composés d'une balise paragraphe, de sauts de ligne et d'un saut de page :

```
<div type="book" n="3">  
  <div type="chapter" n="9">  
    <p>Elisabeth avoit arrangé dans sa tête<lb/> que Mr. Darcy lui amèneroit sa sœur le<lb/> lendemain de son arrivée à Pemberley  
    ,<lb/> et en conséquence elle étoit bien déci-<lb break="no"/>dée à ne pas s'éloigner de l'auberge de<lb/> toute la matinée;  
    mais elle avoit mal<lb/> calculé, car il l'amena à Lambton le<lb/> jour même. Elisabeth et sa tante qui<lb/> s'étoient  
    promenées près de là avec quel-<lb break="no"/>ques-uns de leurs nouveaux amis, ren-<lb break="no"/>troient à l'auberge pour  
    faire leur toi-<lb break="no"/>lette et aller dîner, lorsque le bruit d'un<lb/> équipage les attira vers la fenêtre, et  
    elles<lb/> virent un monsieur et une dame dans un<lb/> Carriole qui s'arrêta à leur porte. Elisabeth,<lb/> reconnoissant la  
    livrée, ne causa pas à son<lb/> oncle et à sa tante une légère surprise,<pb n="113"/><note type="foliation" place="top(right)"  
    >113</note> en leur annonçant la visite qu'elle atten-<lb break="no"/>doit. Jusqu'alors ils n'avoient eu aucun<lb/>  
    soupçon de ce qui se passoit ; mais com-<lb break="no"/>ment expliquer l'embarras d'Elisabeth, et<lb/> toutes les attentions  
    de Darcy? Il falloit<lb/> qu'il eût du penchant pour leur nièce.</p>  
  </div>  
</div>
```



# Le corps du texte : le <body>

## Exemples

### Un poème

```
<div type="poem">
  <head>Exemple de poèmes</head>
  <lg type="sonnet">
    <lg type="quatrain">
      <l>Vers 1</l>
      <l>Vers 2</l>
      <l>Vers 3</l>
      <l>Vers 4</l>
    </lg>
    <lg type="quatrain"> [5 lines]
    <lg type="tercet">
      <l>Vers 1</l>
      <l>Vers 2</l>
      <l>Vers 3</l>
    </lg>
    <lg type="tercet"> [4 lines]
  </lg>
</div>
```

### Un tableau

```
<div type="table">
  <head>Exemple de tableau</head>
  <table>
    <head>Informations personnelles</head>
    <row>
      <cell/>
      <cell>Genre</cell>
      <cell>Profession</cell>
    </row>
    <row>
      <cell>Jean Martin</cell>
      <cell>Homme</cell>
      <cell>Professeur</cell>
    </row>
    <row>
      <cell>Marie Martin</cell>
      <cell>Femme</cell>
      <cell>Chirurgien</cell>
    </row>
  </table>
</div>
```

# Le corps du texte : le <body>

## Exemples

Une figure

```
<div type="figure">
  <head>Exemple de figure</head>
  <figure>
    <graphic url="inr_logo_rouge.jpg" width="2cm"/>
    <head>Logo de l'Inria</head>
    <figDes>L'image montre le logo de l'Inria, écrit
      en rouge et en écriture cursive.</figDes>
  </figure>
</div>
```

Une liste

```
<div type="list">
  <head>Exemple de liste</head>
  <list>
    <head>Liste d'objets</head>
    <item>Une table</item>
    <item>Une chaise</item>
    <item>Un lit</item>
  </list>
</div>
```

# Les éléments spécifiques

3 Elements Available in All TEI Docs <https://tei-c.org/release/doc/tei-p5-doc/fr/html/CO.html#COPU>

13 Names, Dates, People, and Places <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ND.html>

Encodage d'entités nommées, soit tout ce qui a trait aux personnes, lieux et organisations :

## ❑ Encodage générique

- ❑ `<rs>` chaîne de référence
- ❑ `<name>` nom ou nom propre

} S'accompagne d'un attribut `@type` (person, place ou org) et `@ref`

## ❑ Encodage précis

- ❑ `<persName>` nom de personne
- ❑ `<placeName>` nom de lieu
- ❑ `<orgName>` nom d'organisation

} À accompagner d'un attribut `@ref`

# Les éléments spécifiques

Exemple d'utilisation de balises pour entités nommées génériques et spécifiques

```
<div type="named_entities">
  <div type="generic">
    <p><name type="person" ref="#p0001">Elisabeth</name> avait arrangé dans sa tête que <name type="person" ref="#p0002">Mr.
      Darcy</name> lui amènerait <rs type="person" ref="#p0003">sa sœur</rs> le lendemain de son arrivée à <name type="place"
      ref="#l0001">Pemberley</name> , et en conséquence elle étoit bien décidée à ne pas s'éloigner de <rs type="place"
      ref="#l0003">l'auberge</rs> de toute la matinée; mais elle avait mal calculé, car il l'amena à <name type="place"
      ref="#l0002">Lambton</name> le jour même. <name type="person" ref="#p0001">Elisabeth</name> et <rs type="person"
      ref="#p0004">sa tante</rs> qui s'étoient promenées près de là.</p>
  </div>
  <div type="specific">
    <p><persName ref="#p0001">Elisabeth</persName> avait arrangé dans sa tête que <persName ref="#p0002">Mr. Darcy</persName> lui
      amènerait <rs type="person" ref="#p0003">sa sœur</rs> le lendemain de son arrivée à <placeName ref="#l0001"
      >Pemberley</placeName> , et en conséquence elle étoit bien décidée à ne pas s'éloigner de <rs type="place" ref="#l0003"
      >l'auberge</rs> de toute la matinée; mais elle avait mal calculé, car il l'amena à <placeName ref="#l0002"
      >Lambton</placeName> le jour même. <persName ref="#p0001">Elisabeth</persName> et <rs type="person" ref="#p0004">sa
      tante</rs> qui s'étoient promenées près de là.</p>
  </div>
</div>
```

# Construire et documenter son schéma d'encodage (et s'y tenir !)




# Construire son schéma d'encodage

- ❑ Déclarer le jeu de balises de son encodage
- ❑ Quatre types de déclaration de classes TEI :
  - ❑ La classe n'est absolument pas utilisée → pas de déclaration et tous ses éléments seront exclus de l'arbre
  - ❑ La classe est un peu utilisée → déclaration et intégration des éléments utilisés avec @include
  - ❑ La classe est beaucoup utilisée → déclaration et exclusion des éléments non utilisés avec @except
  - ❑ La classe est entièrement utilisée → déclaration et aucune exclusion
- ❑ Possibilité de personnalisation du schéma TEI
  - ❑ Addition/Suppression/Changement d'éléments
  - ❑ Personnalisation d'attributs et de leurs valeurs

# Créer son schéma d'encodage avec Roma

- ❑ Aller sur Roma : <https://romabeta.tei-c.org/>
- ❑ Choisir "TEI ALL" et "START"
- ❑ Mettre comme titre "ODD Formation TEI" et mettre votre nom dans "Author" puis cliquer sur "--> CUSTOMIZE ODD"
- ❑ Classer les éléments par module avec "↗ by module"
- ❑ Désélectionner *"analysis"*, *"certainty"*, *"corpus"*, *"dictionaries"*, *"drama"*, *"gaiji"*, *"iso-fs"*, *"linking"*, *"nets"*, *"spoken"*, *"tagdocs"* et *"textcrit"*
- ❑ Avec l'outil de recherche, taper *"div"* et désélectionner les div numérotés
- ❑ Avec l'outil de recherche, taper *"name"* et désélectionner l'élément *"name"*
- ❑ Avec l'outil de recherche, taper *"del"*, cliquer sur l'élément puis *"Attributes"*, modifier l'attribut *"rend"*, mettre l'option "Closed" et ajouter les valeurs *"erasure"*, *"overwritten"* et *"strikethrough"*
- ❑ Cliquer sur "Download" et choisir "Customization as ODD"
- ❑ Modifier l'extension du fichier pour mettre un ".xml"

# Appliquer son ODD sur un fichier XML-TEI

- ❑ Ouvrir l'ODD nouvellement créée et opérer la transformation en schéma Relax NG (qui est un type de fichier utilisé pour valider la conformité à la TEI)
  - ❑ Appuyer sur le bouton "Appliquer les scénarios de transformation" 
  - ❑ Choisir "TEI ODD to RELAX NG XML"
  - ❑ Appuyer sur le bouton "Appliquer"
- ❑ Dans le dépôt du cours, télécharger le fichier "orgueil\_et\_prejuges.xml" et ensuite l'ouvrir
  - ❑ Appuyer sur le bouton "Associer schéma" 
  - ❑ Dans "URL", utiliser le bouton "Naviguer"  et choisir le fichier RNG nouvellement créé et situé dans le dossier "out"
  - ❑ Dans "Type de schéma", sélectionner "RELAX NG XML Syntax" puis appuyer sur "OK"
- ❑ Relever les erreurs qui se présentent alors et proposer des modifications selon ce qui est nécessaire, pour permettre de rendre le fichier conforme



# Pourquoi documenter son schéma d'encodage ?

## Pourquoi encoder ?

- ❑ Préciser et expliquer les choix faits,
- ❑ Garantir l'homogénéité de son encodage
- ❑ Aider à la réutilisation du type d'encodage fait

## Deux types de documentation :

- ❑ Documentation partielle, présentation des balises dont l'utilisation peut différer par rapport à ce que définit les TEI guidelines
- ❑ Documentation exhaustive, présentation de l'arbre de A à Z, avec description de l'utilisation des éléments/attributs dans l'arbre

# Comment documenter son schéma d'encodage ?

Documentation écrite avec les balises traditionnelles pour l'écriture en prose dans le <body> :

- ❑ <div> : pour la hiérarchie des parties
- ❑ <head> : pour les titres de parties
- ❑ <p> : pour le contenu de la présentation et des descriptions

Balises spécifiques pour les composants de la structure TEI :

- ❑ <gi> : pour les éléments
- ❑ <att> : pour les attributs
- ❑ <val> : pour la valeur des attributs
- ❑ <egXML xmlns="http://www.tei-c.org/ns/Examples"> : pour les exemples tirés de l'arbre XML


# Documenter son schéma d'encodage

## Exercice :

Dans l'ODD que vous avez généré avec *Roma* et avec les informations données précédemment à propos de l'écriture d'une documentation d'ODD, faites un paragraphe qui présente votre choix d'utiliser les balises `<persName>`, `<placeName>` et `<orgName>` pour les entités nommées, plutôt que `<name>` avec un attribut `@type` qui contient les valeurs "person", "place", ou "organisation", et fournissez un exemple tiré du texte de démonstration *Orgueil et Préjugés*.

# Travaux pratiques : définir un premier balisage pour son corpus

# Oxygen XML Editor

- ❑ Éditeur pour tout type de document utilisant le format XML : <https://www.oxygenxml.com/>
- ❑ Autocomplétion → si *namespace* déclaré et vocabulaire défini, suggestions d'utilisation de balises, selon la hiérarchie imposée par la TEI
- ❑ Quelques petits raccourcis pour l'utilisation d'Oxygen
  - ❑ Balisage rapide : Cmd + E (Mac) ; Ctrl + E (Linux)
  - ❑ Indenter : Bouton  ou Cmd + Shift + P (Mac) ; Ctrl + Shift + P (Linux)
  - ❑ Recherche/Remplace : Cmd + F (Mac) ; Ctrl + F (Linux)
- ❑ Quelques expressions régulières utiles pour l'utilisation du Recherche/Remplace
  - ❑ ^ → Début de ligne | \$ → Fin de ligne
  - ❑ [0-9] → Nombre entre 0 et 9 | [a-zA-Z] → Caractère entre a et z
  - ❑ [^...] → Exclusion de caractères
  - ❑ ? → 0 ou 1 occurrence | + → 1 ou plusieurs occurrences

# Encoder les métadonnées

## **Exercice :**

Créer un fichier XML-TEI que vous enregistrerez sous le nom de “Le\_tour\_du\_monde\_en\_80\_jours.xml” dans un endroit facile d’accès (sur le Bureau ou dans les Téléchargements par exemple)

En utilisant les métadonnées de l’ouvrage *Le Tour du monde en quatre-vingts jours* présent dans le dossier “texte\_de\_travail” du dépôt du cours et sur la page Gallica de l’ouvrage, ainsi qu’en vous aidant de ce que vous pourrez dans les [TEI Guidelines](#) et dans les textes d’exemple, encoder le <teiHeader> de l’ouvrage utilisé comme texte de travail.

# Encoder le corps du texte

## Exercice :

En utilisant le fichier texte présent dans le dossier “texte\_de\_travail” du dépôt du cours, ainsi que les ressources que vous pouvez trouver dans les [TEI Guidelines](#), encoder le <body> du chapitre 2 du *Tour du monde en quatre-vingts jours* de Jules Verne, afin de représenter tous les éléments que l’on peut trouver dans cet extrait de l’ouvrage (numéro de chapitre, titres, pagination, figures, etc.)

Astuce pour encoder rapidement les linebreaks (après avoir encodé les paragraphes, titres, pagination, etc.):

- ❑ Sélectionnez votre texte et faites chercher/remplacer
- ❑ Vérifier que “Seulement les lignes sélectionnées” et “Expressions régulières” sont cochés
- ❑ Dans “Rechercher”, rentrez “-\$” puis faites “Chercher tout” pour vérifier ce que l’on vous propose
- ❑ Une fois que vous êtes sûrs, dans “Remplacer”, rentrez “-<lb break=“no”/>” et faites “Tout remplacer”
- ❑ Refaites l’expérience en changeant “-\$” par “[^>]\$" et “<lb/> ” (ne fonctionnera que si les fins des lignes sont précédés d’espace, sinon cela remplacera un caractère dans les mots de fin de ligne, d’où la nécessité de vérifier ce qui va être remplacé)

# Encoder les entités nommées

## **Exercice:**

En reprenant votre texte nouvellement encodé et en vous aidant de ce qui a précédemment été cité en exemple, ainsi que des [TEI Guidelines](#), encoder les noms des deux personnages principaux du chapitre (pour chacune de leurs mentions), tous les lieux que vous trouverez, ainsi que l'organisation qui est mentionnée à quelques reprises.

Vous devrez ensuite créer un index dans le header, où vous regrouperez, par type, les différentes entités que vous aurez trouvées.



# Documentation



# Ressources

- ❑ TEI guidelines :  
<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>
- ❑ Oxygen XML Editor :  
<https://www.oxygenxml.com>
- ❑ Roma : <https://romabeta.tei-c.org/>
- ❑ Documentation ODD :  
<https://tei-c.org/release/doc/tei-p5-doc/en/html/TD.html>