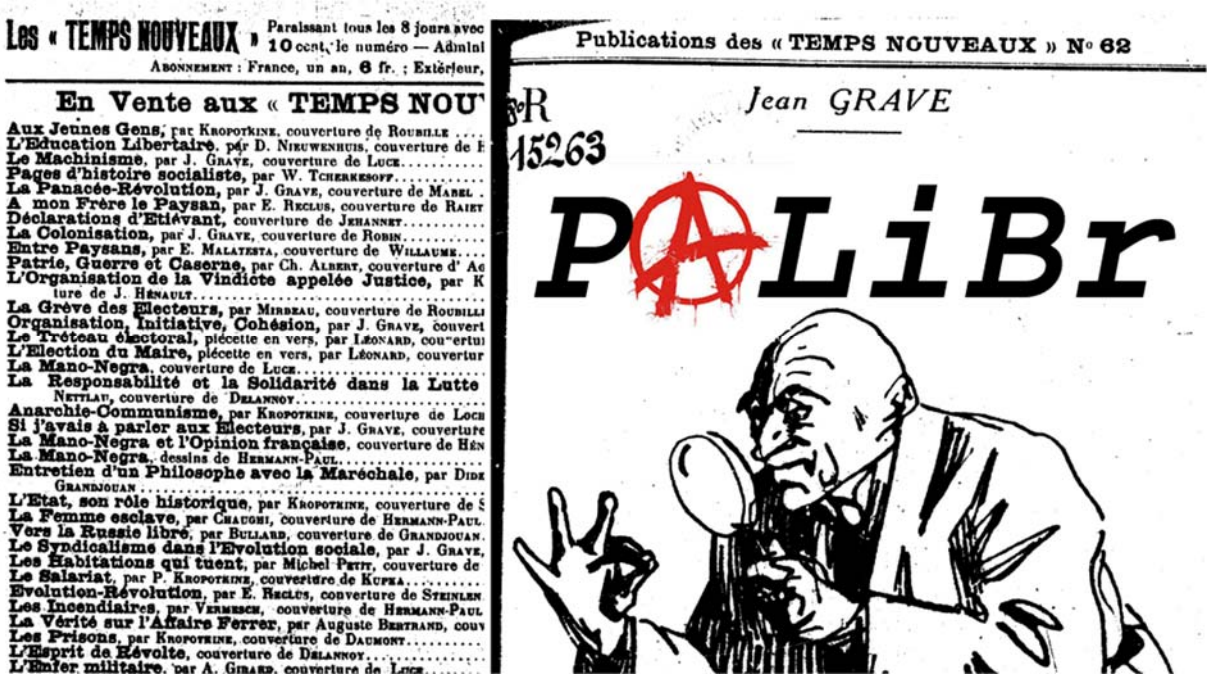
Encodage TEI projet PALiBR



Recommandations pour l'application de la TEI au corpus des brochures anarchistes PALiBr
Barbara Bonazzi
Version du 29/11/2022

Table des matières

Mené au sein du CODHOS (Collectif des centres de documentation en histoire ouvrière et sociale) par le CHS depuis plusieurs années, le projet PALiBr, porté par le Grand équipement documentaire du Campus Condorcet et par le CHS, vise à mettre à la disposition de la communauté scientifique un corpus numérisé et océrisé d'environ 800 brochures anarchistes francophones publiées de 1880 à 1918 en France, mais également en Suisse ou en Belgique, terres d'accueil d'anarchistes en exil. Ce corpus dispersé au sein de différents établissements partenaires du CODHOS sera consultable sur la

plateforme dédiée aux sciences humaines et sociales, celle de la bibliothèque numérique du Campus Condorcet. Au-delà de l'intérêt patrimonial, ce corpus virtuel offrira une vue exhaustive de la production anarchiste de la période permettant d'explorer de nouvelles pistes de recherche, en croisant les points de vue disciplinaires.

Outre qu'exposées dans la bibliothèque numérique de l'Humathèque Condorcet, les fichiers produits par la numérisation pourront être exploités dans le cadre de projets d'édition numérique en TEI, comme celui imaginé par le CELLF et le CHS qui n'a pas pu voir le jour pour le moment. Ce projet prévoyait l'affichage simultané de plusieurs versions de la même brochure, qu'il s'agisse de nouvelles éditions ou d'autres tirages, pour étudier l'évolution des textes.

Le projet PALiBr a été mis à profit pour permettre la montée en compétences des partenaires concernés, en particulier autour de l'encodage en XML-TEI. A long terme, l'encodage pourra servir à créer une édition numérique où l'on pourra visualiser, cote à cote ou superposées, les différentes versions et/ou éditions d'un même titre.

Comme le prévoyait le projet, Claude Rétat du CELLF et Barbara Bonazzi du CHS ont pu bénéficier d'une formation opérationnelle sur la TEI suivie d'une expérimentation à partir des brochures numérisées. Ainsi, elles ont été formées à la TEI et au logiciel Oxygen par Julie Aucange de l'IR HumaNum Loire/MSH Ange Guépin de Nantes.

Le schéma et le modèle TEI présentés dans le présent document sont le résultat de la formation amélioré à l'aide de Raphaëlle Lapotre, Chargée de soutien à la Science Ouverte de l'Humathèque Condorcet.

Ce document a été rédigé sur la base des Guidelines de la TEI en ligne sur http://www.tei-c.org/ et des supports de formation de Julie Aucange. Il est complété, en annexe, par une liste des éléments et attributs utilisés.

Pour la compréhension du concept de balisage et de l'encodage, voir le par exemple le manuel d'utilisation des Bibliothèques Virtuelles des Humanistes sur https://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/Manuel_TEI_BVH.html

Établi entre juin et novembre 2022, ce document ne doit pas être tenu comme le résultat de l'encodage du projet PALiBr : il représente une première réflexion sur la possible utilisation de la TEI dans l'encodage des brochures numérisées pour créer des index utiles dans les outils de recherche.

Introduction

La TEI est un langage de balisage fondé sur le XML (eXtensible Markup Langage) : langage de balisage, de structuration sémantique des documents.

L'XML permet de structurer des données textuelles sans perte de contenu (contrairement à un format tableau, bdd…). Il sépare les données (le texte) des métadonnées (les balises) : celles-ci sont lues par la machine et peuvent être manipulées à différentes fins.

Les principes du balisage

Ce métalangage est caractérisé par des **balises** ouvrantes et fermantes `<texte></texte>`, **imbriquées** `<oeuvre> <titre>` Ce que nous voulons `</titre></oeuvre>`

Le nombre de balises est illimité et l'XML ne dépend d'aucune technologie particulière et peut donc être écrit et déployé sur de nombreux supports.

Il comprend au minimum :

- La déclaration : `<?xml version="1.0" encoding="UTF-8"?>`
- La « racine » du document, qui contient l'arborescence : `<TEI xmlns="http://www.tei-c.org/ns/1.0">` La balise est fermée à la fin du document `</TEI>`

Éléments, attributs et valeurs :

Un **élément** est contenu dans une balise ouvrante et fermante :

```
<nom> PALiBr </nom>
```

Il peut être qualifié par un **attribut** (ou plusieurs, séparés par un espace):

```
<titre lang=''fr'' > Ce que nous voulons </titre>
```

Auquel on associe une **valeur** (ex. =''fr''), placée entre guillemets

Il est possible d'associer à son document un **schéma personnalisé** (= une grammaire) qui contraint l'usage de ces éléments (lesquels sont autorisés, à quelles conditions, etc.)

La TEI

La TEI adopte toutes les règles syntaxiques propres au XML et y superpose les siennes, qui sont d'ordre surtout sémantique, en définissant des éléments, en imposant des règles d'imbrication, etc.

- La TEI propose des règles partagées par une communauté de recherche
- Les données encodées en TEI sont donc interopérables et réutilisables dans d'autres contextes

La TEI est utilisée :

- Pour mettre à disposition des ressources de toute nature sous forme de **données numériques structurées** (la TEI est surtout centrée sur le texte mais permet aussi l'encodage d'autres types de ressources)

Et en particulier pour

- Établir des éditions savantes
- Générer des index, thesauri, dictionnaires…
- Étudier, par des moyens informatiques, certains aspects d'un corpus donné

Un document TEI doit avoir une

```
<teiHeader>Obligatoire : il s'agit des métadonnées du fichier XML/TEI
     <fileDesc>Obligatoire : Description bibliographique de la ressource et
     de ses origines
            <titleStmt> contient l'ensemble des informations sur le titre
           du document numérique et ceux qui l'ont créé.
                  <title> titre du document numérique</title>
                  <author> auteur du document numérique (si document
                  natif)</author>
                  <editor> responsabilité secondaire du document
            numérique</editor>
                  <respStmt> responsabilité intellectuelle (=
                  collaborateurs sur le fichier numérique, comme le
                  transcripteur, l'encodeur etc) </respStmt>
                  <name> ou <persName> nom de la personne concernée</name>
                  ou </persName>
                  <resp> fonction</resp>
                  <date> date </date>
            </titleStmt>
            <publicationStmt> Obligatoire : contient les informations sur
             les lieux et l'éditeur de la ressource numérique.
                  <publisher> nom de l'éditeur ou du responsable de la
                  publication numérique</publisher>
                  <pubPlace> (publication place) : lieu
            d'édition</pubPlace>
                  <date> date d'édition de la ressource numérique</date>
                  <availability> décrit les droits de la ressource
            numérique</availability>
                  <p> Information sur la publication du fichier </p>
            </publicationStmt>
            <sourceDesc> Obligatoire : décrit la source originale du
            document numérique. S'il n'y a pas d'original et que le
            document est un document numérique "natif", on le précise ici.
                  <p>Information sur la source du fichier numérique</p>
                  <p>On peut aussi introduire ici des listes pour générer
                  des index … </p>

                  <listPerson> de personnes</listPerson>
                  <listPlaces>de lieux géographiques</listPlaces>
                  <listOrg>de collectivités </listOrg>
            </sourceDesc>
      </fileDesc>
</teiHeader>

<text>Obligatoire :
      <front>facultatif</front>
      <body>obligatoire</body>
      <back> facultatif </back>
</text>
```

Les choix retenus pour le corpus PALiBr

Dans le cadre de cette phase d'encodage TEI du corpus PALiBr, nous avons choisi de créer :

- un fichier « Corpus » (corpus_PALiBr.xml), un seul document TEI maître. Celui-ci contiendra les textes des brochures à la suite les uns des autres ainsi qu'une description du corpus, sa

genèse, ses objectifs, les choix opérés, etc. Cela permettra d'éviter des redondances dans le <teiHeader> de chaque document, puisque ces métadonnées sont communes à tous les documents du corpus. Toutes les brochures encodées seront collées, les unes après les autres et matérialisées par la présence de balises <TEI> successives ;

- un fichier « Modèle » (modele_tei_Palibr.xml) de balisage qui sert à encoder chaque brochure du corpus. On a décidé de produire des métadonnées de qualité dans l'en-tête de l'encodage de la brochure (balise <teiHeader>), en créant des index de noms de personnes et d'organisations, ainsi que des lieux géographiques, présents dans le texte (via des balises de listes de personnes, d'organisation ou encore de lieux) ;
- un exemple de brochure encodée (CC_PALiBr_exemple.xml) ;
- un schéma TEI

Fichier « Corpus » : la balise <teiCorpus>

Pour encoder plusieurs documents TEI appartenant à un même corpus, il convient d'ajouter la balise `<teiCorpus>` qui décrira le corpus et chaque texte le composant.

L'élément `<teiCorpus/>` aura son propre `<teiHeader/>` et comprendra tous les documents encodés en `<TEI/>`, chacun avec son propre `<teiHeader/>` et son `<text/>`.

Exemple :

```
<teiCorpus version="5.2" xmlns="http://www.tei-c.org/ns/1.0">
      <teiHeader><!-- header du corpus -->
      <fileDesc>
            <titleStmt>
            <title>PALiBr La pensée libertaire par les brochures</title>
                  <author>CODHOS Collectif des centres de documentation en
                histoire ouvrière et sociale</author>
            </titleStmt>
      </fileDesc>
      </teiHeader>
      <texte><! – texte de présentation du corpus  -- ></texte>
      <TEI>
      <teiHeader><!—header de la première brochure encodée -->
            </teiHeader>
      <text><!—contenu de la première brochure --></text>
</TEI>
 <TEI>
      <teiHeader><!-- header de la deuxième brochure encodée -->
      </teiHeader>
      <text><!-- contenu de la deuxième brochure --></text>
 </TEI>
<TEI><!—d'autres éléments TEI à la suite -->

      </TEI>

</teiCorpus>
```

Fichier « Modèle » : l'indexation

La possibilité de créer des index est l'un des principaux intérêts d'un encodage en TEI car l'index permet d'uniformiser les éléments sujets à des variations, tels que les noms, les lieux, etc.

La réalisation d'un ou plusieurs index est conseillé pour réaliser un outil de recherche performant.

Dans le cadre du modèle PALiBr il a été décidé de créer les index dans le `<sourceDesc>` de la `<teiHeader>`.

Les index crées concernent :

- **Les individus**, où l'élément <listPerson> peut contenir de véritables notices biographiques et constituer un élément majeur de l'appareil critique

```
<listPerson>
            <person xml:id="Per1">
                    <persName ref="url dans un référenciel, ex. IdREF, le
                    dictionnaire Maitron">
                       <forename>Prénom</forename>
                       <surname>Nom</surname>
                    </persName>
           </person>
                 <person xml:id="Per2">
                     <persName ref="url dans un référenciel, ex. IdREF, le
                    dictionnaire Maitron">
                          <forename>Prénom</forename>
                          <surname>Nom</surname>

                    </persName>
           </person>
      …
   </listPerson>
```

- **Les lieux cités** qui peuvent aussi faire l'objet de description très précises

```
<listPlace>
          <head>Liste des lieux cités</head>
          <place>
                  <placeName xml:id="Geo1"> Nom géographique</placeName>
                  <geo>coordonnées géographiques</geo>
          </place>
          <place>
                   <placeName xml:id="Geo2"> Nom géographique </placeName>
          </place>
                  …
  </listPlace>
```

- **Les éditeurs et maisons d'édition**

```
<listOrg>
         <org>
              <orgName xml:id="Edit1">Premier éditeur</orgName>
         </org>
         <org>
               <orgName xml:id="Edit2">autre éditeur</orgName>
         </org>
      …
</listOrg>
```

- **Les imprimeurs**

```
<listOrg>
         <org>
              <orgName xml:id="Imp1">Imp. Coop. La Laborieuse
               </orgName>
          </org>
          <org>
```

```
            <orgName xml:id="Imp2"/>
         </org>
      …
</listOrg>
```

Le texte

L'élément <text/> englobe l'ensemble du texte.

```
<text>
      <front> tout ce qui est au début du document </front>
      <body> le corps du document </body>
        <div> pour séparer des chapitres, des parties, des textes
        différents </div>
           <p>texte</p>
      <back> tout ce qui est mis après le texte principal </back>
</text>
```

Nous avons décidé de baliser plus finement la page de titre et de ne baliser que les paragraphes pour le corps du texte.

Exemple :

- **La page de titre**

```
<text>
      <front n="couverture">
         <p>20c</p>
         <p>
            <persName ref="#Per1">
                  <forename>Pierre</forename>
                  <surname>CHARDON</surname>
            </persName>
          </p>
         <p>
            <title>CE QU'EST LA PATRIE</title>
          </p>
         <p>(portrait-bois de <persName ref="#Per2">L. Moreau</persName>)
        </p>
         <p><persName ref="#Per3">E. ARMAND</persName></p>
         <p>
             <title>Le Refus de Service Militaire et sa véritable
                    signification</title>
          </p>
         <p>-^LJWTtQr<!-- illustration --></p>
         <p> Le soldat idéal : pas de tête et tout en muscles</p>
         <p>
            <orgName ref="#Edit1">Editions de l'en dehors</orgName>
          </p>
         <p>22, cité Saint-Joseph, Orléans</p>
         <p>DEUXIÈME TIRAGE</p>
         <p>
            <placeName ref="#Geo1">Orléans</placeName> — <orgName
                        ref="#Imp1">Imp. Coop. La Laborieuse</orgName>.
         </p>
      </front>
</text>
```

- **Le corps du document**

```
<body>
    <div type="texte" n="1">
        <head>
                <title>Ce qu'est la Patrie</title>
                <ref target="https://bibliothèque  numérique du Campus
                 Condorcet"/>
        </head>
        <p>
           <persName>
                <forename>Pierre</forename>
                 <surname>Chardon</surname>
           </persName>
         </p>
        <p>Novembre 1892-Mai 1919</p>
        <p>Le mot Patrie est un dérivé, une variante, une déformation du
           mot Pater, c'est-à-dire père, tout comme patriarche, patricien,
           pa-trice ou patron.
              ...</p>
           <p>…</p>
      </div>
      <div type="texte" n="2">
         <head>
           <title>Le Refus de Service Militaire et sa véritable
           signification (Rapport présenté au Congrès Antimilitariste
           International d'Amsterdam)</title>
         </head>
          <p>Au mois de juin 1904 se tint à Amsterdam un Congrès
             Antimilitariste International auquel je pris part ...</p>
          <p>  20 février 1925.
            <persName ref="#Per3"><forename>
             E.</forename><surname>Armand</surname></persName>,condamné le
             5 janvier 1918 à cinq années d'emprisonnement par le Conseil
             de Guerre de Grenoble, sous le prétexte d'avoir a favorisé-la
             désertion d'un militaire.»</p>
            <p>….</p>
      </div>
</body>
```

- **La fin**

```
<back>
         <p>Publicité </p>
</back>
```

Le schéma XML

Produire son propre schéma, adapté à son projet spécifique, et le contraindre au maximum permet d'éviter les erreurs et les incohérences dans l'encodage (surtout s'il y a plusieurs collaborateurs), pour le projet PALiBr nous avons créé un schéma à partir du modèle TEI Lite dans Roma: generating customizations for the TEI.

Annexe Liste des balises utilisées

Liste des balises utilisées établie à partir des Guidelines for Electronic Text Encoding and Interchange

- <author> (author) in a bibliographic reference, contains the name(s) of an author, personal or corporate, of a work; for example in the same form as that provided by a recognized bibliographic name authority.
- <availability> (availability) supplies information about the availability of a text, for example any restrictions on its use or distribution, its copyright status, any licence applying to it, etc.
- <back> (back matter) contains any appendixes, etc. following the main part of a text.
- <body> (text body) contains the whole body of a single unitary text, excluding any front or back matter.
- <date> (date) contains a date in any format.
- (text division) contains a subdivision of the front, body, or back of a text.
- <fileDesc> (file description) contains a full bibliographic description of an electronic file
- <forename> (forename) contains a forename, given or baptismal name.
- <front> (front matter) contains any prefatory matter (headers, abstracts, title page, prefaces, dedications, etc.) found at the start of a document, before the main body.
- <head> (heading) contains any type of heading, for example the title of a section, or the heading of a list, glossary, manuscript description, etc.
- <list> (list) contains any sequence of items organized as a list.
- <listOrg> (list of organizations) contains a list of elements, each of which provides information about an identifiable organization.
- <listPerson> (list of persons) contains a list of descriptions, each of which provides information about an identifiable person or a group of people, for example the participants in a language interaction, or the people referred to in a historical source.
- <listPlace> (list of places) contains a list of places, optionally followed by a list of relationships (other than containment) defined amongst them.
- <org> (organization) provides information about an identifiable organization such as a business, a tribe, or any other grouping of people.
- <orgName> (organization name) contains an organizational name.
- <p> (paragraph) marks paragraphs in prose.
- <person> (person) provides information about an identifiable individual, for example a participant in a language interaction, or a person referred to in a historical source.
- <persName> (personal name) contains a proper noun or proper-noun phrase referring to a person, possibly including one or more of the person's forenames, surnames, honorifics, added names, etc.
- <place> (place) contains data about a geographic location
- <placeName> (place name) contains an absolute or relative place name.
- <pubPlace> (publication place) contains the name of the place where a bibliographic item was published.
- <publicationStmt> (publication statement) groups information concerning the publication or distribution of an electronic or other text.
- <publisher> publisher) provides the name of the organization responsible for the publication or distribution of a bibliographic item.
- <sourceDesc> (source description) describes the source(s) from which an electronic text was derived or generated, typically a bibliographic description in the case of a digitized text, or a phrase such as "born digital" for a text which has no previous existence.

- <surname> (surname) contains a family (inherited) name, as opposed to a given, baptismal, or nick name.
- <TEI> (TEI document) contains a single TEI-conformant document, combining a single TEI header with one or more members of the model.resource class. Multiple TEI elements may be combined within a TEI (or teiCorpus) element.
- <teiCorpus/> contains the whole of a TEI encoded corpus, comprising a single corpus header and one or more TEI elements, each containing a single text header and a text
- <teiHeader> supplies descriptive and declarative metadata associated with a digital resource or set of resources.
- <text> (text) contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, a dictionary, or a corpus sample.
- <title> (title) contains a title for any kind of work.
- <titleStmt> (title statement) groups information about the title of a work and those responsible for its content.