

Atelier XML TEI

Ariane Pinche ¹

¹CNRS, ²CIHAM ■ UMR 5648

30 novembre 2023, Université d'Avignon

Table of Contents

- 1 Qu'est-ce que XML TEI
 - XML
 - TEI
 - Structurer un document TEI
- 2 Éléments de base en TEI
- 3 Les TEIguidelines
- 4 Structurer un index

XML est un format de données pur, très simple et documenté, conçu pour la **description** des documents textuels. XML ne possède pas de jeu de balises prédéfini.

```
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

XML : un standard international

Depuis 1998, XML est un langage libre et documenté. XML est également un **langage standard** respectant les recommandations du **W3C** (World Wide Web Consortium), il facilite :

- La lisibilité par les machines ou par l'œil humain
- L'échange de données
- La migration vers d'autres plates-formes, d'autres logiciels, d'autres formats

XML est né en 1996 sous la tutelle du W3C (World Wide Web Consortium).

- SGML (1970), Standard Generalized Markup Language;
 - ▶ HTML, HyperText Markup Language: affiche des données notamment sur le Web;
 - ▶ XML, eXtensible Markup Language: contient et structure des données textuelles.

XML s'inscrit dans un environnement complet et possède des langages associés qui permettent de parser, de transformer et d'interroger les fichiers XML.

- Xpath est un langage pour naviguer dans les documents XML :
TEI/text/body/lg/l
- XSLT est un langage de transformation pour les documents XML
- Xquery est un langage pour interroger les documents XML

Les éléments

- Les données sont incluses dans le document XML sous forme de chaînes de caractères délimitées par un balisage les décrivant
- L'unité de base qui comprend données et balisage est appelée élément
Exemple : `<nomElement>chaineCaracteres</nomElement>`
- Les éléments peuvent être vides : `<element>texte</element>` ou `<elementVide/>`
- Les éléments XML suivent un principe strict d'arborescence par imbrication.
- Les éléments *enfants* héritent des propriétés des éléments *parents*.

Les attributs

- Les attributs XML peuvent être multipliés autant que nécessaire.
- On ne peut pas ajouter deux fois le même attribut sur un élément.
- Dans un attribut, on peut mettre plusieurs valeurs.

```
<MiseEnValeur rendu='rouge gras' position='centre'>  
texte  
</MiseEnValeur>
```

Quelques règles importantes :

- À chaque balise de début doit correspondre une fin de balise.
- Les éléments peuvent être imbriqués, mais ils ne doivent pas se recouvrir.
- Il ne doit y avoir qu'un seul élément racine.
- Un élément ne doit pas avoir deux attributs avec le même nom.

Un encodage qui respecte ces grands principes est **bien formé**.

Bien formé ou pas ?

`<paragraphe>du texte</paragraphe>`

`<paragraphe><article>du</article><nom>texte</nom></paragraphe>`

`<paragraphe><article>du <nom></article>texte</nom></paragraphe>`

`<paragraphe type="texte">du texte</paragraphe>`

`<paragraphe type=texte>du texte</paragraphe>`

`<paragraphe type="texte">du texte<paragraphe/>`

`<paragraphe type="texte">du texte<nomPersonnage>nom de personnage</paragraphe>`

`<paragraphe type="texte">du texte</Paragraphe>`

`<segment type="texte" type="nombre">du texte</paragraphe>`

Quels sont les avantages de TEI ?

- XML TEI permet de proposer un vocabulaire commun pour les balises
- XML TEI s'intéresse au sens du texte plutôt qu'à son apparence
- XML TEI est indépendant de tout environnement logiciel
- XML TEI a été conçu par la communauté scientifique qui est aussi en charge de son développement continu

- 1987 : établissement de la *Text Encoding Initiative*
- 1990 : TEI P1 (proposal 1), dir. Michael Sperberg-McQueen et Lou Burnard
- 1994 : TEI P3, première version complète;
- 2000 : naissance du TEI Consortium
- 2001-2004 : TEI P4, introduction du XML
- 2007-... : TEI P5, abandon de SGML

La communauté TEI est animée par le **TEI consortium**, fondation interdisciplinaire à but non lucratif.

Il se compose des unités suivantes:

- TEI Board of Directors
- TEI Technical Council
- Membres institutionnels et individuels
- TEI Workgroups, par exemple :
 - ▶ TEI Manuscripts Special Interest Group
 - ▶ Correspondence SIG

La communauté peut échanger et se rencontrer grâce à :

- Une liste de diffusion internationale : TEI-L mailing list;
- Une liste francophone : TEI-FR et un wiki;
- Des congrès annuels : TEI Conference;
- Une revue : Journal of the Text Encoding Initiative;
- Des Guidelines ("recommandations") qui documentent notamment chaque élément.

TEI est un set de balises prédéfini et documenté dans les TEIguidelines qui permet de procéder à une description scientifique et **sémantique** d'un texte.

Les balises TEI forment un framework, utile à la conception de son propre encodage. **Il est fortement déconseillé d'utiliser l'intégralité de la TEI pour un document.** Il faut concevoir un modèle de données le plus simple possible et adapté à son projet et sa question de recherche.

Pour aller plus loin : *Why do we encode* : E. Pierazzo

Table of Contents

- 1 Qu'est-ce que XML TEI
 - XML
 - TEI
 - Structurer un document TEI
- 2 Éléments de base en TEI
- 3 Les TEIguidelines
- 4 Structurer un index

Structuration générale d'un fichier XML TEI

Tout document TEI a au moins deux parties :

- Un en-tête, représenté au moyen d'un élément `<teiHeader>` contenant des métadonnées décrivant le document ;
- Le texte lui-même, représenté par un élément `<text>` qui peut être subdivisé en sous-unités.

```
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
  <teiHeader>
    <!-- métadonnées décrivant un texte -->
  </teiHeader>
  <text>
    <!-- texte -->
  </text>
</TEI>
```

Structuration du `teiHeader`

Le **teiHeader** minimal comporte au minimum un élément *fileDesc* contenant les trois sections suivantes :

- un élément **titleStmt** : informations identifiant le document lui-même
- un élément **publicationStmt** : informations sur la façon dont il est distribué ou publié
- un élément **sourceDesc** : indications sur ses origines

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Titre de l'œuvre</title>
    </titleStmt>
    <publicationStmt>
      <p>Informations sur la publication de l'œuvre.</p>
    </publicationStmt>
    <sourceDesc>
      <p>Informations sur la source du texte</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Structuration du corps du texte

l'élément **text** peut contenir les éléments suivants :

- un élément **front** pour le texte liminaire (avant-propos, préface, etc..)
- un élément **body** pour le corps de texte
- un élément **back** pour les annexes, postface, etc.

```
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
  <teiHeader>
    <!-- métadonnées décrivant un texte -->
  </teiHeader>
  <text>
    <front/>
    <body/>
    <back/>
  </text>
</TEI>
```

Principaux éléments de structuration du texte

L'information textuelle peut être hiérarchisée au moyen des éléments suivants

- **div** pour les grandes divisions ou subdivisions du texte
- **p** pour les paragraphes
- **lg** pour les strophes
- **l** pour les vers
- **sp** pour les répliques.
- Ces éléments peuvent être personnalisés avec les attributs *type* et *n*

```
<text>
  <body>
    <div type="book" n="1">
      <div type="chapter" n="1">
        <p>Some text here.</p>
      <p/>
    </div>
  </div>
</body>
</text>
```

Structurer en XML TEI le poème de Paul Verlaine, *Mon rêve Familier*

Lien vers le texte : https://fr.wikipedia.org/wiki/Mon_rêve_familier

- Ouvrir Oxygen Editor
- Ouvrir un nouveau fichier *XML TEI all*
- Remplir un `teiHeader` minimal
- Copier le texte de Verlaine dans le **body**
- Structurer le poème en vous aidant des éléments vus précédemment

Table of Contents

- 1 Qu'est-ce que XML TEI
 - XML
 - TEI
 - Structurer un document TEI
- 2 Éléments de base en TEI
- 3 Les TEIguidelines
- 4 Structurer un index

Les recommandations visent à :

- Fournir un format standard
- Favoriser l'échange de textes
- Proposer des ensembles de conventions d'encodage adaptés à des applications différentes

TEI guidelines, mode d'emploi

Comment lire les guidelines ?

La page d'accueil :

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

TEI < Text Encoding Initiative >

P5 Guidelines — English Search

P5: Guidelines for Electronic Text Encoding and Interchange

Version 3.2.0. Last updated on 10th July 2017, revision 0fc651

[\[English\]](#) [\[Deutsch\]](#) [\[Español\]](#) [\[Italiano\]](#) [\[Français\]](#) [\[日本語\]](#) [\[한국어\]](#) [\[中文\]](#)



Front Matter

[Title](#)

- i. [Releases of the TEI Guidelines](#)
- ii. [Dedication](#)
- iii. [Preface and Acknowledgments](#)
- iv. [About These Guidelines](#)
- v. [A Gentle Introduction to XML](#)
- vi. [Languages and Character Sets](#)

Back Matter

- Appendix A [Model Classes](#)
- Appendix B [Attribute Classes](#)
- Appendix C [Elements](#)
- Appendix D [Attributes](#)
- Appendix E [Datatypes and Other Macros](#)
- Appendix F [Bibliography](#)
- Appendix G [Prefatory Notes](#)
- Appendix H [Colophon](#)

Text Body

- 1 [The TEI Infrastructure](#)
- 2 [The TEI Header](#)
- 3 [Elements Available in All TEI Documents](#)
- 4 [Default Text Structure](#)
- 5 [Characters, Glyphs, and Writing Modes](#)
- 6 [Verse](#)
- 7 [Performance Texts](#)
- 8 [Transcriptions of Speech](#)
- 9 [Dictionaries](#)
- 10 [Manuscript Description](#)
- 11 [Representation of Primary Sources](#)
- 12 [Critical Apparatus](#)
- 13 [Names, Dates, People, and Places](#)
- 14 [Tables, Formulae, Graphics and Notated Music](#)
- 15 [Language Corpora](#)
- 16 [Linking, Segmentation, and Alignment](#)
- 17 [Simple Analytic Mechanisms](#)
- 18 [Feature Structures](#)
- 19 [Graphs, Networks, and Trees](#)
- 20 [Non-hierarchical Structures](#)
- 21 [Certainty, Precision, and Responsibility](#)
- 22 [Documentation Elements](#)
- 23 [Using the TEI](#)

TEI sourcecode

- [Getting and Using the TEI Sources](#),
- [TEI GitHub Repository](#)
- [Bug Reports, Feature Requests, etc.](#)

Comment lire les spécifications d'élément ?



<msIdentifier>

Accueil
C Éléments

<msIdentifier> (identifiant du manuscrit) Contient les informations requises pour identifier le manuscrit en cours de description. [\[10.4 The Manuscript Identifier\]](#)

Module	msdescription — Manuscript Description	L'élément est documenté dans le module msdescription (10 Manuscript Description)
Attributs	<code>att_global @xml:id, @n, @xmi:lang, @xmi:base, @xmi:space (att_global.rendition (@rend, @style, @rendition)) (att_global.linking (@corresp, @synch, @sameAs, @copyOf, @next, @prev, @exclude, @useless)) (att_global.analytic (@ana)) (att_global.faces (@facs)) (att_global.change (@change)) (att_global.responsibility (@cert, @resp)) (att_global.source (@source))</code>	attribut class
Membre du	model.bibliPart	model class
Contenu dans	core: bibl msdescription: msDesc msFrag msPart	éléments regroupés par modules
Peut contenir	header: idno msdescription: altIdentifier collection institution msName repository namesdates: bloc country district geoName placeName region settlement	
Exemple	<pre><msIdentifier> <country>France</country> <settlement>Paris</settlement> <repository xml:lang="fr">Bibliothèque nationale de France. Réserve des livres rares</repository> <idno>= 73</idno> <!-- dans le cas des recueils : cote uniquement sans les sous-cotes --> <altIdentifier> <idno>= 121</idno> <note> Cote de la bibliothèque royale au XVIIIe siècle (inscrite à l'encre, sur la doublure de table).</note> </altIdentifier> <altIdentifier> <idno>=Double de B. 274. A (Réserve)</idno> <note>Cote inscrite face à la page de titre, en remplacement de la cote "1541", barrée</note> </altIdentifier> </msIdentifier></pre>	Toute la liste bibliography

En vous aidant des TEI guidelines :

- Trouver la documentation pour encoder des répliques de pièce de théâtre
- Comment peut-on indiquer qui prononce la réplique ?

Convertir en XML TEI, un extrait du Misanthrope de Molière en vous aidant de la documentation et des balises suivantes : `<castList>`, `<castItem>`, `<speaker>`, `<stage>`, `<head>`, `<div>`, `<sp>` et `<l>`.

Lien vers le texte :

https://fr.wikisource.org/wiki/Le_Misanthrope/Édition_Louandre,_1910

Documentation dans les guidelines :

<https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html>

Table of Contents

- 1 Qu'est-ce que XML TEI
 - XML
 - TEI
 - Structurer un document TEI
- 2 Éléments de base en TEI
- 3 Les TEIguidelines
- 4 Structurer un index

Structuration d'un index

Structuration des index en XML TEI

Les entités nommées sont signalées à l'aide des balises *persName* ou *placeName*.

```
<text n="edition">
  <body n="urn:cts:froLit:jns915.jns2117.ciham-fro1">
    <div n="1"
      type="chapter"
      facs="http://gallica.bnf.fr/ark:/12148/btv1b84259980/f237, http://gallica.bnf.fr/ark:/12148/btv1b84259980/f238"
      corresp="urn:cts:latinLit:stoa0270.stoa003:1.1, urn:cts:latinLit:stoa0270.stoa003:1.2">
      <head>Ci coumence</b></li dialogues que <persName ref="#postumien">postumiens</persName> <persName ref="#gallus">Gal
      <persName ref="#severus">seuerus</persName> le
      rate</b></ les oi <app>
        <lem>ler</lem>
        <rdg wit="#C2 #C3" type="ajout">parler et qi i fu</rdg>
      </app>
    </head>
    <div n="1" type="section">
      <p>
        <pb n="114" facs="http://gallica.bnf.fr/ark:/12148/btv1b84259980/f237"/>
        <hi rend="rubricated orig">Ci coumence</b></li dialogues que <persName ref="#postumien">postumiens</persName>
        de <persName ref="#martin">seint martin</persName> des
        moines d<placeName ref="#egypte" full="yes">egy</b></b>pte</placeName> si <persName ref="#severus">seuerus</persName>
        oi ler</hi>
        <lb/>
        <hi rend="decorated-initial">U</hi>N ior auint</b></b>qeie
      <persName ref="#gallus">gallus</persName>
        <lb/>
        <cb n="b"/>
      Et mes chiers compains qi deciples Fu</b></b>seint martin estions ensamble laou
      nos</b></b> lions depluiseurs choses<persName ref="#postumien">Postumiens</persName>.i.</b></b>miens tres chiers amisuint
      <placeName ref="#orient" full="yes">orit</placeName>
        <lb/>deuers la terre d<placeName full="yes">egypte</placeName>.ouil auoit<seg type="number">iii</seg>
    </p>
  </div>
</body>
</text>
```

Structuration d'un index

Pour que toutes les occurrences d'une entité puissent pointer vers une entrée unique, il faut la déclarer soit dans l'élément *profilDesc* du *teiHeader* soit dans un élément *standOff* dans une liste de personnes ou de lieux.

```
<listPerson>
  <person xml:id="agaridut">
    <persName>Agariduz</persName>
    <note type="biographical">Honnête homme, témoin de l'adoucissement des mœurs cruelles
d'Avicien.</note>
  </person>
  <person xml:id="agnes">
    <persName>Agnies</persName>
    <death when="0300">4<ex>e</ex> siècle</death>
    <note type="biographical" source="http://data.bnf.fr/11958468/agnes/">Martyre
romaine.</note>
  </person>
  <person xml:id="antoine">
    <persName>Antoines</persName>
    <birth when="0251">251</birth>
    <death when="0356">356</death>
    <note type="biographical" source="http://data.bnf.fr/11888967/antoine/">
<surname type="complex" full="yes">Antoine le grand, Antoine d'Égypte ou Antoine
l'Ermite</surname>, considéré comme le fondateur du l'érémittisme chrétien.</note>
  </person>
  <person xml:id="arborius">
    <persName>Arborius</persName>
    <note type="biographical">Magnus Arborius, élu comte des largesses sacrées en 379, puis
préfet de Rome en 380, il aurait été l'un des protecteurs de <persName ref="#saintMartin">saint Martin</persName>.</note>
  </person>
  <person xml:id="arpage">
    <persName>Arpage</persName>
    <note type="biographical">Prêtre.</note>
  </person>
```

Documentation:

<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html>

Lien entre l'index et les occurrences grâce à un jeu d'identifiants et de pointeurs.

- L'identifiant est unique, il est déclaré dans la liste des entités de l'index avec l'attribut *xml:id*
- ex : `<person xml:id="agaridut">`
- Le pointeur est ajouté sur les occurrences de l'entité dans le texte avec l'attribut *ref*
- ex : `<persName ref="#agaridut">`
- La valeur du pointeur est la chaîne de caractères de @xml:id précédée d'un dièse.

Encoder les deux premiers paragraphes de *Notre Dame de Paris*, livre II, chapitre 1, «De Charybde en Scylla » (début de chapitre à « La cohue admirait »).

- Structurer le texte en livre, chapitre et paragraphes
- Baliser les noms de lieux
- Réaliser un index

Lien vers le texte : https://fr.wikisource.org/wiki/Notre-Dame_de_Paris/Livre_deuxième#II._La_place_de_Gr.C3.A8ve

- Téléchargez le texte produit lors de l'atelier HTR
- Structurez votre texte en XML
- Remplir un `teiHeader` minimal
- Pour encoder une source manuscrite voir la documentation suivante : [Representation of Primary Sources](#)
- Pour structurer une lettre voir la documentation suivante : [Default Text Structure](#)