UNIL | Université de Lausanne

Faculté des lettres

Centre NUCLEUS

Formation de deux jours

Initiation à l’encodage XML-TEI

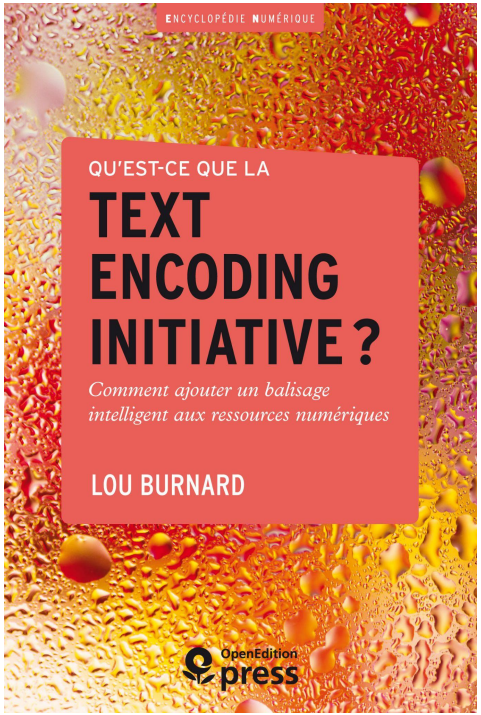
Lausanne, février 2025

Sonia Solfrini et Simon Gabay

(Université de Genève)

Plan du cours 1 : introduction

- Qu'est-ce que la TEI ?
- La syntaxe du langage XML
- La structure minimale d'un document XML-TEI
- L'éditeur XML-Oxygen
- Quelques exercices pratiques
- Les TEI guidelines



La « *Text Encoding Initiative* » (TEI) est l'un des projets les plus durables et influents du champ aujourd'hui appelé « humanités numériques ». Son but est de fournir des recommandations pour la création et la gestion sous forme numérique de tout type de données créées et utilisées par les chercheurs en sciences humaines, comme les sources historiques, les manuscrits, les documents d'archives, les inscriptions anciennes et bien d'autres.

https://books.openedition.org/oep/1237

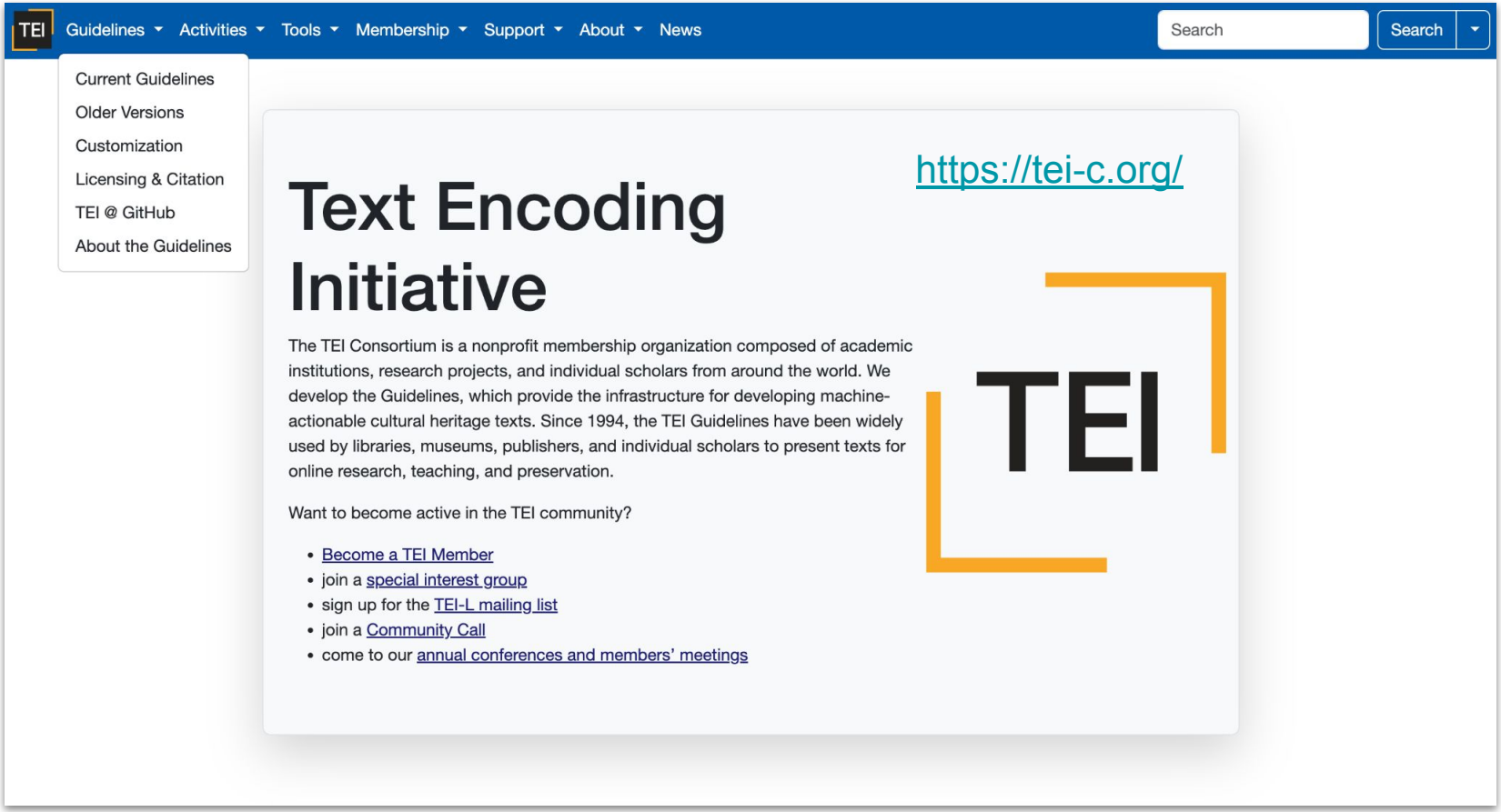
XML-TEI ?

Les documents TEI reposent sur la syntaxe du langage XML. Un document TEI est donc dit bien formé s'il respecte la syntaxe de XML, avec des balises ouvrantes et fermantes correctement imbriquées.

Le document TEI est dit valide si les balises qu'il contient se conforment à un schéma qui, en plus des règles syntaxiques de XML, fournit des recommandations sur le vocabulaire à utiliser. Les *Guidelines* de la TEI fournissent le nom et la définition de centaines de balises, ainsi que des règles sur la façon dont elles peuvent être combinées.

N.B. : La plupart des documents TEI n'a besoin que d'une petite partie de ce qui est fourni !

N.B.B. : On n'invente pas de balises et elles sont en anglais !


TEI
Guidelines
Activities
Tools
Membership
Support
About
News
Search
Search
Current Guidelines
Older Versions
Customization
Licensing & Citation
TEI @ GitHub
About the Guidelines
https://tei-c.org/
Text Encoding Initiative
The TEI Consortium is a nonprofit membership organization composed of academic institutions, research projects, and individual scholars from around the world. We develop the Guidelines, which provide the infrastructure for developing machine-actionable cultural heritage texts. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation.
Want to become active in the TEI community?
Become a TEI Member
join a special interest group
sign up for the TEI-L mailing list
join a Community Call
come to our annual conferences and members' meetings
TEI

XML

XML ou « *eXtensible Markup Language* », publié pour la première fois en 1998 par le World Wide Web Consortium (W3C), est un langage d'encodage formel très largement utilisé.

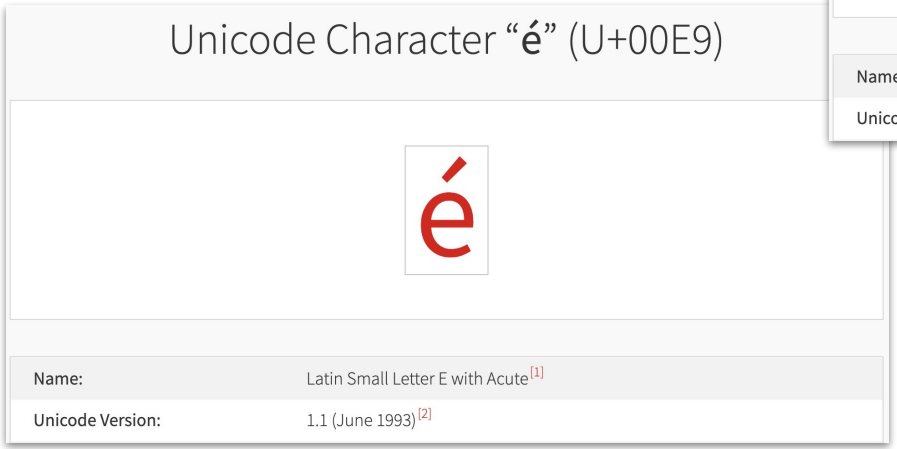
La première ligne d'un document XML consiste toujours en une déclaration indiquant que ce qui suit est un document XML conforme à la version du standard XML indiquée :

```
<?xml version="1.0" encoding="UTF-8"?>
```

De plus, dans la déclaration XML, on peut indiquer le système d'encodage des caractères utilisé. UTF-8 est l'un des encodages les plus utilisés car il est capable de représenter n'importe quel caractère Unicode.

Petit rappel :

exemples de caractères Unicode



La syntaxe du langage XML

Les balises (en anglais « *tags* ») sont les unités de base dans le langage XML et elles sont marquées par des chevrons « < » et « > ». Elles sont utilisées pour définir les éléments et leur contenu.

Un élément est composé d'une balise ouvrante et d'une balise une fermante : <élément>blabla</élément>.

Exemple : <sentence>Bienvenue à ce cours d'initiation à l'encodage XML-TEI !</sentence>

N.B : Une balise peut être auto-fermante quand un élément est vide. Dans ce cas, attention à la position de la barre oblique qui se trouve à droite : <élément/>.

Exemple : <pb/> (page break)

La syntaxe du langage XML

N.B. Les éléments ne se chevauchent jamais, ils sont imbriquées les uns dans les autres.

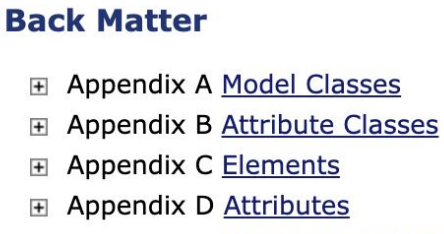
Exemple : <sentence>Bienvenue à ce cours d'<italique>initiation à l'encodage XML-TEI</italique> !</sentence>

La syntaxe du langage XML

Un élément peut porter un ou plusieurs attributs qui le décrivent avec une valeur pour chaque attribut : <élément attribut="valeur">blabla</élément>

Exemple : <date when="2009-01-20">20 janvier 2009</date>

Exemple : <sentence type="citation" who="Obama">Yes, we can!</sentence>



La structure minimale d’un document TEI

Un fichier XML-TEI est organisé comme une arborescence d’éléments et les relations sont hiérarchiques.

Voici la structure minimale d’un document TEI :

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <!---…-->
    </teiHeader>
    <text>
        <!--…-->
    </text>
</TEI>
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Titre du texte électronique.</title>
      </titleStmt>
      <publicationStmt>
        <p>Informations concernant la publication ou la distribution d'un texte électronique.
          Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </publicationStmt>
      <sourceDesc>
        <p>La ou les sources à partir desquelles un texte électronique a été dérivé ou généré.
          Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Contenu du texte électronique.
        Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
    </body>
  </text>
</TEI>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Titre du texte électronique.</title>
         </titleStmt>
         <publicationStmt>
            <p>Informations concernant la publication ou la distribution d'un texte électronique.
               Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
         </publicationStmt>
         <sourceDesc>
            <p>La ou les sources à partir desquelles un texte électronique a été dérivé ou généré.
               Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <p>Contenu du texte électronique.
            Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </body>
   </text>
</TEI>
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Titre du texte électronique.</title>
      </titleStmt>
      <publicationStmt>
        <p>Informations concernant la publication ou la distribution d'un texte électronique.
          Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </publicationStmt>
      <sourceDesc>
        <p>La ou les sources à partir desquelles un texte électronique a été dérivé ou généré.
          Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Contenu du texte électronique.
        Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
    </body>
  </text>
</TEI>
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Titre du texte électronique.</title>
         </titleStmt>
         <publicationStmt>
            <p>Informations concernant la publication ou la distribution d'un texte électronique.
               Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
         </publicationStmt>
         <sourceDesc>
            <p>La ou les sources à partir desquelles un texte électronique a été dérivé ou généré.
               Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <p>Contenu du texte électronique.
            Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </body>
   </text>
</TEI>
```

L’éditeur XML-Oxygen

Oxygen XML est un outil pour la création et le développement XML. Il offre des fonctionnalités avancées pour éditer et publier du contenu XML.

Quelques atouts d’Oxygen :

- Quand vous commencez à écrire une balise, il suggère les éléments possibles.
- Il ferme automatiquement les éléments que vous venez d’ouvrir.
- Le carré vert/rouge indique si votre document est valide, c’est-à-dire s’il respecte la syntaxe XML et le schéma utilisé (TEI guidelines).

Structure_minimale.xml

TEI teiHeader

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Titre du texte électronique.</title>
            </titleStmt>
            <publicationStmt>
                <p>Informations concernant la publication ou la distribution d'un texte électronique.</p>
            </publicationStmt>
            <sourceDesc>
                <p>La ou les sources à partir desquelles un texte électronique a été dérivé ou généré.</p>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <p>Contenu du texte électronique.</p>
        </body>
    </text>
</TEI>
```

Structure_minimale.xml*

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>

    </teiHeader>
    <text>
        <body>
            <p>Contenu du texte électronique.</p>
        </body>
    </text>
</TEI>
```

L’éditeur XML-Oxygen

Pour utiliser Oxygen XML vous avez besoin d’une licence, payante ou fournie par votre institution. Il est également possible d’obtenir une licence d’essai de 30 jours.

Plan du cours 1 : introduction

- Qu'est-ce que la TEI ?
- La syntaxe du langage XML
- La structure minimale d'un document XML-TEI
- L'éditeur XML-Oxygen
- Quelques exercices pratiques
- Les TEI guidelines

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Titre du texte électronique.</title>
      </titleStmt>
      <publicationStmt>
        <p>Informations concernant la publication ou la distribution d'un texte électronique.
          Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </publicationStmt>
      <sourceDesc>
        <p>La ou les sources à partir desquelles un texte électronique a été dérivé ou généré.
          Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Contenu du texte électronique.
        Pas forcément dans la balise "p", mais au moins une balise est attendue.</p>
    </body>
  </text>
</TEI>
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Définition de TEI.</title>
      </titleStmt>
      <publicationStmt>
        <p>Exemple pour un cours d'initiation à l'encodage XML-TEI.</p>
      </publicationStmt>
      <sourceDesc>
        <p>Burnard, Lou. Qu'est-ce que la Text Encoding Initiative ?. Traduit par Marjorie Burghart, OpenEdition Press, 2015, https://doi.org/10.4000/books.oep.1237.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <pb/>
      <p>La « Text Encoding Initiative » (TEI) est l'un des projets les plus durables et influents du champ aujourd'hui
        appelé « humanités numériques ». Son but est de fournir des recommandations pour la création et la
        gestion sous forme numérique de tout type de données créées et utilisées par les chercheurs en sciences
        humaines, comme les sources historiques, les manuscrits, les documents d'archives, les inscriptions
        anciennes et bien d'autres.
      </p>
    </body>
  </text>
</TEI>
```

XML-TEI ? Révisons !

Les documents TEI reposent sur la syntaxe du langage XML. Un document TEI est donc dit bien formé s'il respecte la syntaxe de XML, avec des balises ouvrantes et fermantes correctement imbriquées.

Le document TEI est dit valide si les balises qu'il contient se conforment à un schéma qui, en plus des règles syntaxiques de XML, fournit des recommandations sur le vocabulaire à utiliser. Les *Guidelines* de la TEI fournissent le nom et la définition de centaines de balises, ainsi que des règles sur la façon dont elles peuvent être combinées.

N.B. : La plupart des documents TEI n'a besoin que d'une petite partie de ce qui est fourni !

N.B.B. : On n'invente pas de balises et elles sont en anglais !

Plan du cours 1 : introduction

- Qu'est-ce que la TEI ?
- La syntaxe du langage XML
- La structure minimale d'un document XML-TEI
- L'éditeur XML-Oxygen
- Quelques exercices pratiques
- Les TEI guidelines

TEI: Guidelines for Electronic Text Encoding and Interchange

P5 Version 4.8.1. Last updated on 1st November 2024, revision 0a2bff95a

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]

Front Matter

Back Matter

Text Body

TEI sourcecode

- Getting and Using the TEI Sources.
- TEI GitHub Repository
- Bug Reports, Feature Requests, etc.

https://tei-c.org/