Introduction au langage XML pour l'analyse de texte

une série d'ateliers

Les mardis de 10 h 30 à 12 h (FR)
Les mardis de 13 h 30 à 15 h (EN)

Avec Alice Defours et
Constance Crompton



Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML
    a. Exercice
3. Utiliser les guidelines de la TEI
4. Encodade : Niveau 1
5. Encodage : Niveau 2
6. Encodage : Niveau 3



Atelier III

31 mars 2026

1. Rappels importants



Rappel : **La TEI : Comment ça fonctionne ?**

Grands éléments : Livres - Chapitres
Moyens éléments : Paragraphes - Strophes
Petits éléments : Phrases - Mots

- **Grands éléments**

= les principales divisions du document :
** ** : segmentation structurelle de base, pouvant s'imbriquer récursivement (elle signifie « division ». Vous pouvez avoir autant de divisions imbriquées que nécessaire pour votre texte).

- **Eléments Moyens**

Éléments structurels de niveau intermédiaire pouvant s'insérer dans une division principale et pouvant contenir contenant des petits éléments.
<head>: <head>The Wonderful Wizard of Oz**</head>**
<p>: paragraphes
<list>: listes (<list> doit contenir <items> pour être valide)
<item>: contenu par <list>

- **Petits éléments**

Petits éléments représentant des segments de la taille d'une phrase ou d'un mot :
persName: <persName>Baron Olivier of Brighton**</persName>**
placeName: <placeName>New York**</placeName>**
orgName: <orgName>Podunk Sewing Club**</orgName>**
title: <title>Hamlet**</title>, <title>**La Gioconda**</title>**

Rappel

Les éléments non-spatiaux représentent des objets absents :

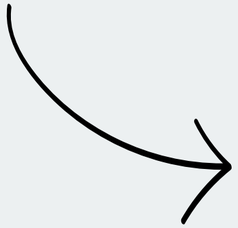
- **<lb/>** : Début de ligne
- **<pb/>** : Début de page
- **<cb/>** : Début de colonne
- **<anchor/>** : Point non spécifié

On en parlera plus tard. Pour l'instant, retenez simplement qu'il existe des éléments représentant des objets textuels qui n'occupent aucun espace physique.



Eléments non-spatiaux

Eléments génériques



Eléments génériques :

Avec plus de 500 éléments, la TEI contient déjà surement tout ce dont vous avez besoin pour décrire votre texte, mais au cas où du cas contraire, vous pouvez utiliser des éléments génériques :

- ** ** division : balise générique pour les grands blocs de texte
- **<ab>** bloc anonyme : pour les blocs de texte de taille moyenne ou de paragraphe
- **<seg>** segment : pour les divisions de la taille d'une phrase ou d'un mot
- **<anchor/>** ancre : pour marquer un point sur la page

Vous pouvez ajouter des attributs et des valeurs pour mieux décrire l'élément :

Exemple : **<ab** type="step"**>**1. Préchauffer le four à 200 °C**</ab>**

Rappel :

Elements obligatoires pour la validité du doc

Structure :
1 <TEI> et son namespace
1 <teiHeader>
1 <text>
1 <body>

Enveloppements

Hiérarchie :
certains éléments ne peuvent pas être enveloppés par d'autres, ou cela cause une erreur

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
3     <teiHeader>
4        <!-- Plus de trucs sur l'en-tête le jour 3. -->
5        <fileDesc>
6           <titleStmt>
7              <title/>
8           </titleStmt>
9           <publicationStmt>
10             <ab/>
11          </publicationStmt>
12          <sourceDesc>
13             <ab/>
14          </sourceDesc>
15       </fileDesc>
16
17       <profileDesc/>
18
19    </teiHeader>
20    <text>
21       <body>
22          <div type="book">
23             <head n="1">LA VÉNUS D'ILLE</head>
24
25             <p>Je descendais le dernier coteau du Canigou, et, bien que le soleil fût déjà
26                couché, je distinguais dans la plaine les maisons de la petite ville d'Ille,
27                vers laquelle je me dirigeais.</p>
28          <p><said> Vous savez</said> dis-je au Catalan qui me servait de guide depuis la veille</p>
29          <p><said>vous savez sans doute où demeure M. de Peyrehorade?</said></p>
30          <p><said>Si je le sais !</said> s'écria-t-il,</p>
31          <p><said>je connais sa maison comme la mienne; et s'il ne faisait pas si noir, je
32                vous la montrerais. C'est la plus belle d'Ille. Il a de l'argent, oui, M. de
33                Peyrehorade; et il marie son fils à plus riche que lui encore.</said></p>
34
35          </div>
36       </body>
37    </text>
38 </TEI>
```

Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML



2. Naviguer sur l’arbre XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
    <teiHeader>
        <!-- Plus de trucs sur l'en-tête le jour 3. -->
        <fileDesc>
            <titleStmt>
                <title/>
            </titleStmt>
            <publicationStmt>
                <ab/>
            </publicationStmt>
            <sourceDesc>
                <ab/>
            </sourceDesc>
        </fileDesc>

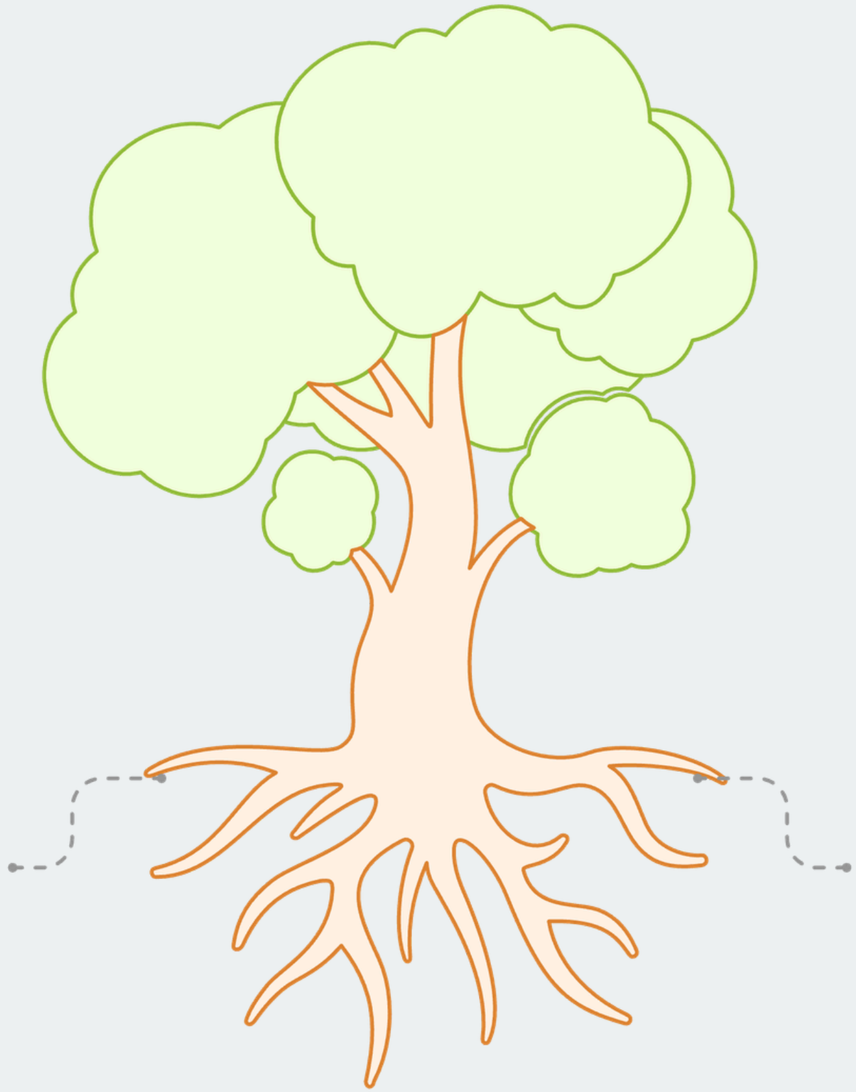
        <profileDesc/>

    </teiHeader>
    <text>
        <body>
            <div type="book">
                <head n="1">LA VÉNUS D’ILLE</head>

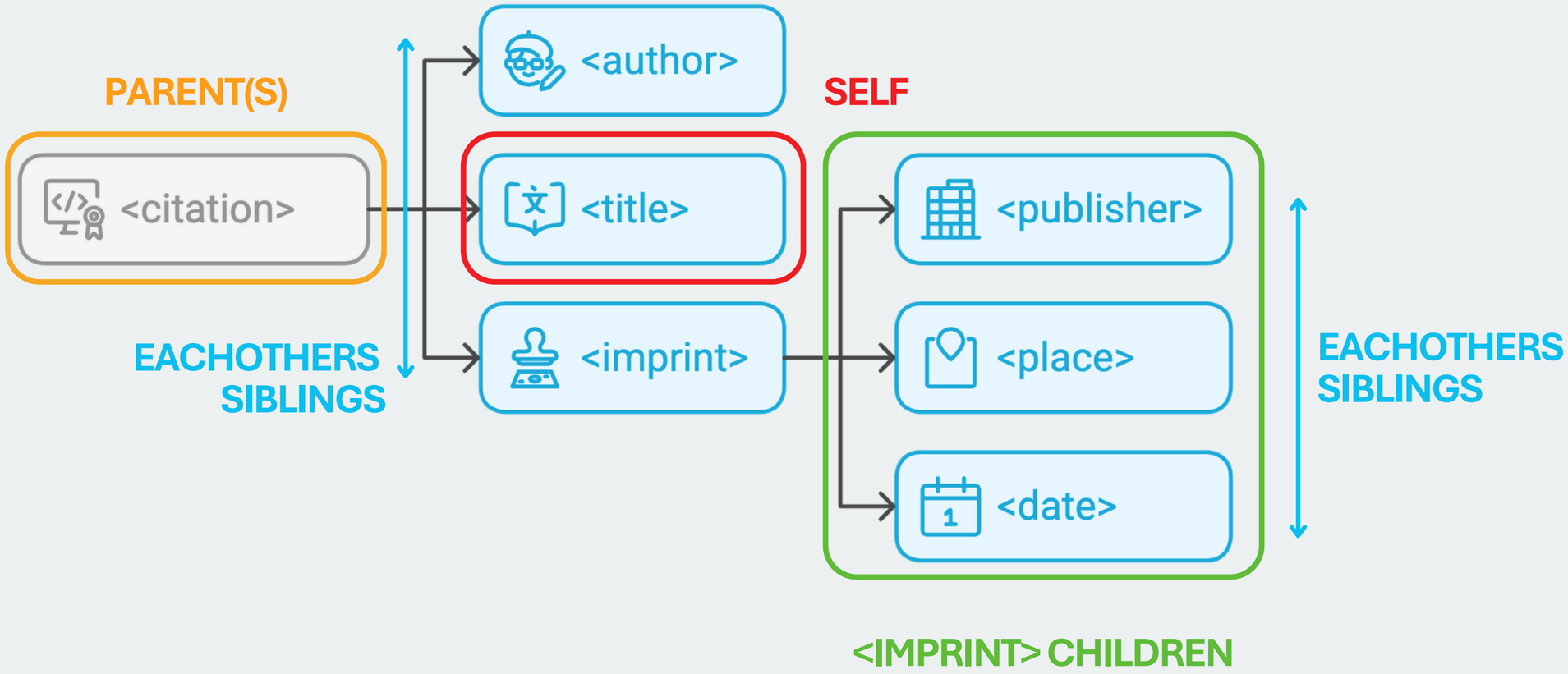
                <p>Je descendais le dernier coteau du Canigou, et, bien que le soleil fût déjà
                    couché, je distinguais dans la plaine les maisons de la petite ville d’Ille,
                    vers laquelle je me dirigeais.</p>
                <p><said> Vous savez</said> dis-je au Catalan qui me servait de guide depuis la veille</p>
                <p><said>vous savez sans doute où demeure M. de Peyrehorade?</said></p>
                <p><said>Si je le sais !</said> s’écria-t-il,</p>
                <p><said>je connais sa maison comme la mienne; et s’il ne faisait pas si noir, je
                        vous la montrerais. C’est la plus belle d’Ille. Il a de l’argent, oui, M. de
                        Peyrehorade; et il marie son fils à plus riche que lui encore.</said></p>

            </div>
        </body>
    </text>
</TEI>
```

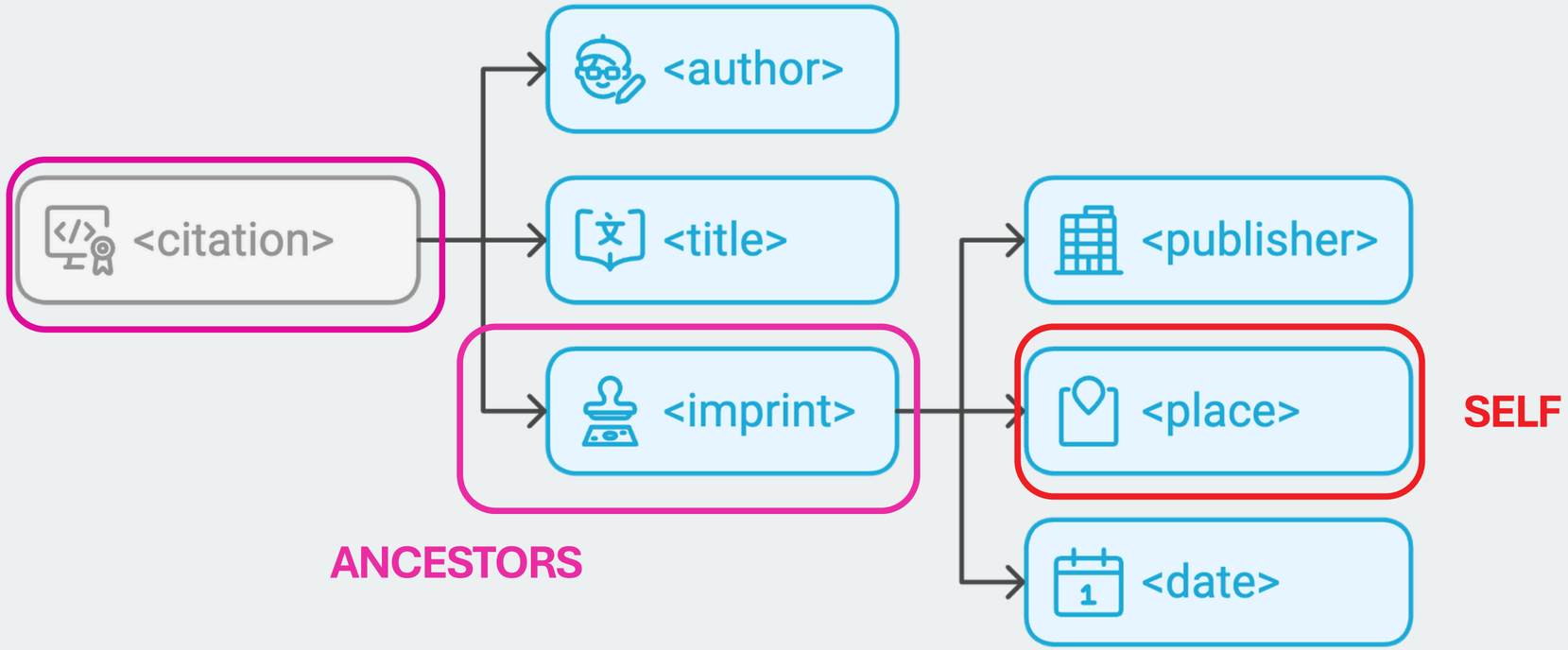
=



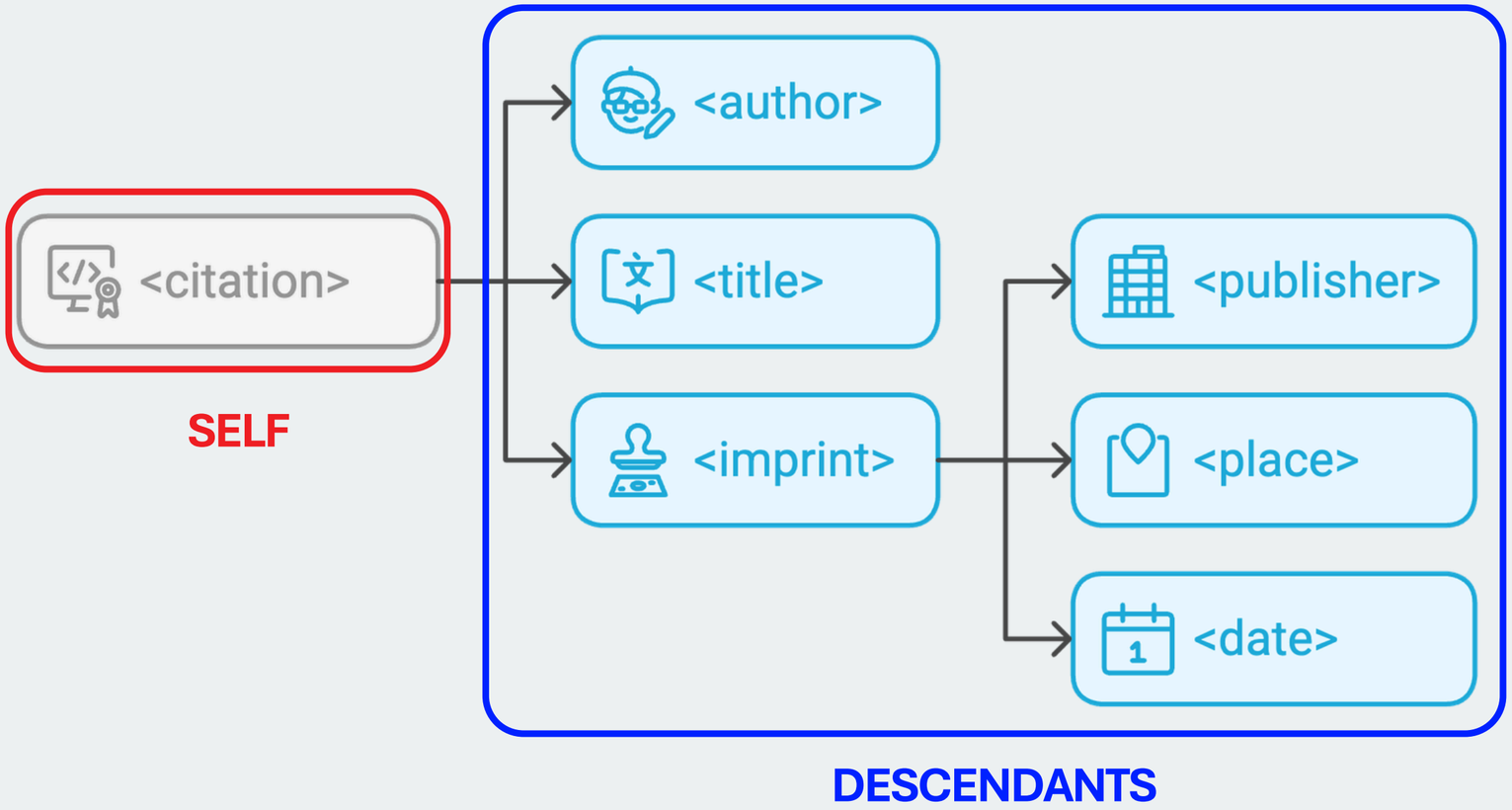
2. Naviguer sur l’arbre XML



2. Naviguer sur l’arbre XML



2. Naviguer sur l’arbre XML



Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML
    a. Exercice



2. Naviguer sur l'arbre XML

Si le deuxième <p> dans la est **SELF**, alors qui sont ses **PARENT(S)**, **SIBLINGS**, **CHILD(REN)**, **ANCESTOR(S)**, **DESCENDANT(S)** ?

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
    <teiHeader>
        <!-- Plus de trucs sur l'en-tête le jour 3. -->
        <fileDesc>
            <titleStmt>
                <title/>
            </titleStmt>
            <publicationStmt>
                <ab/>
            </publicationStmt>
            <sourceDesc>
                <ab/>
            </sourceDesc>
        </fileDesc>

        <profileDesc/>

    </teiHeader>
    <text>
        <body>
            <div type="book">
                <head n="1">LA VÉNUS D'ILLE</head>

                <p>Je descendais le dernier coteau du Canigou, et, bien que le soleil fût déjà
                    couché, je distinguais dans la plaine les maisons de la petite ville d'Ille,
                    vers laquelle je me dirigeais.</p>
                <p><said> Vous savez</said> dis-je au Catalan qui me servait de guide depuis la veille</p>
                <p><said>vous savez sans doute où demeure M. de Peyrehorade?</said></p>
                <p><said>Si je le sais !</said> s'écria-t-il,</p>
                <p><said>je connais sa maison comme la mienne; et s'il ne faisait pas si noir, je
                        vous la montrerais. C'est la plus belle d'Ille. Il a de l'argent, oui, M. de
                        Peyrehorade; et il marie son fils à plus riche que lui encore.</said></p>

            </div>
        </body>
    </text>
</TEI>
```

2. Naviguer sur l'arbre XML

ANCESTORS

PARENT(S)

EACHOTHERS
SIBLINGS

CHILD
SELF

DESCENDANTS ?

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
    <teiHeader>
        <!-- Plus de trucs sur l'en-tête le jour 3. -->
        <fileDesc>
            <titleStmt>
                <title/>
            </titleStmt>
            <publicationStmt>
                <ab/>
            </publicationStmt>
            <sourceDesc>
                <ab/>
            </sourceDesc>
        </fileDesc>

        <profileDesc/>

    </teiHeader>
    <text>
        <body>
            <div type="book">
                <head n="1">LA VÉNUS D'ILLE</head>

                <p>Je descendais le dernier coteau du Canigou, et, bien que le soleil fût déjà
                    couché, je distinguais dans la plaine les maisons de la petite ville d'Ille,
                    vers laquelle je me dirigeais.</p>
                <p><said> Vous savez</said> dis-je au Catalan qui me servait de guide depuis la veille</p>
                <p><said>vous savez sans doute où demeure M. de Peyrehorade?</said></p>
                <p><said>Si je le sais !</said> s'écria-t-il,</p>
                <p><said>je connais sa maison comme la mienne; et s'il ne faisait pas si noir, je
                        vous la montrerais. C'est la plus belle d'Ille. Il a de l'argent, oui, M. de
                        Peyrehorade; et il marie son fils à plus riche que lui encore.</said></p>

            </div>
        </body>
    </text>
</TEI>
```

2. Naviguer sur l'arbre XML

ANCESTORS

PARENT(S)

EACHOTHERS
SIBLINGS

CHILD
SELF

DESCENDANTS ?

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
    <teiHeader>
        <!-- Plus de trucs sur l'en-tête le jour 3. -->
        <fileDesc>
            <titleStmt>
                <title/>
            </titleStmt>
            <publicationStmt>
                <ab/>
            </publicationStmt>
            <sourceDesc>
                <ab/>
            </sourceDesc>
        </fileDesc>

        <profileDesc/>

    </teiHeader>
    <text>
        <body>
            <div type="book">
                <head n="1">LA VÉNUS D'ILLE</head>

                <p>Je descendais le dernier coteau du Canigou, et, bien que le soleil fût déjà
                    couché, je distinguais dans la plaine les maisons de la petite ville d'Ille,
                    vers laquelle je me dirigeais.</p>
                <p><said> Vous savez</said> dis-je au Catalan qui me servait de guide depuis la veille</p>
                <p><said>vous savez sans doute où demeure M. de Peyrehorade?</said></p>
                <p><said>Si je le sais !</said> s'écria-t-il,</p>
                <p><said>je connais sa maison comme la mienne; et s'il ne faisait pas si noir, je
                        vous la montrerais. C'est la plus belle d'Ille. Il a de l'argent, oui, M. de
                        Peyrehorade; et il marie son fils à plus riche que lui encore.</said></p>

            </div>
        </body>
    </text>
</TEI>
```

2. Naviguer sur l'arbre XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
    <teiHeader>
        <!-- Plus de trucs sur l'en-tête le jour 3. -->
        <fileDesc>
            <titleStmt>
                <title/>
            </titleStmt>
            <publicationStmt>
                <ab/>
            </publicationStmt>
            <sourceDesc>
                <ab/>
            </sourceDesc>
        </fileDesc>

        <profileDesc/>

    </teiHeader>
    <text>
        <body>
            <div type="book">
                <head n="1">LA VÉNUS D'ILLE</head>

                <p>Je descendais le dernier coteau du Canigou, et, bien que le soleil fût déjà
                    couché, je distinguais dans la plaine les maisons de la petite ville d'Ille,
                    vers laquelle je me dirigeais.</p>
                <p><said> Vous savez</said> dis-je au Catalan qui me servait de guide depuis la veille</p>
                <p><said>vous savez sans doute où demeure M. de Peyrehorade?</said></p>
                <p><said>Si je le sais !</said> s'écria-t-il,</p>
                <p><said>je connais sa maison comme la mienne; et s'il ne faisait pas si noir, je
                        vous la montrerais. C'est la plus belle d'Ille. Il a de l'argent, oui, M. de
                        Peyrehorade; et il marie son fils à plus riche que lui encore.</said></p>

            </div>
        </body>
    </text>
</TEI>
```

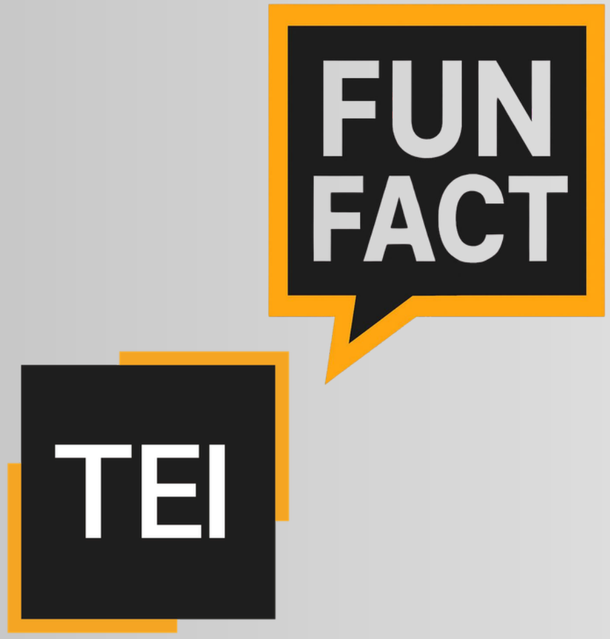
DESCENDANTS

Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML
    a. Exercice
3. Utiliser les Guidelines de la TEI





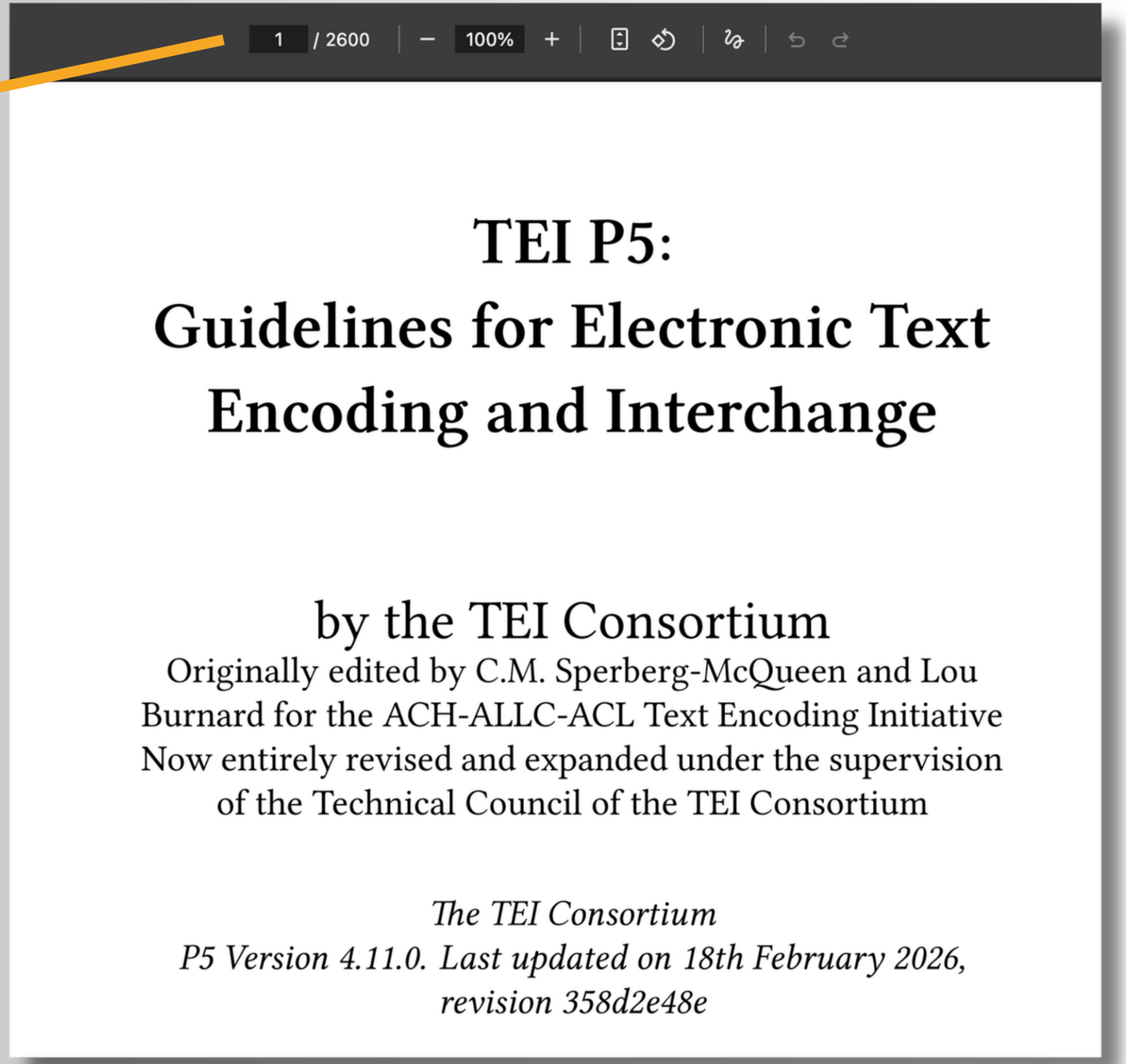
A votre avis, quelle taille font les TEI Guidelines ?
(en pages, si on parle du fichier PDF)

1 / 2600

A votre avis, quelle taille font les TEI Guidelines ?
(en pages, si on parle du fichier PDF)

FUN FACT

TEI





La TEI possède plus de 500 éléments !

Allez consulter les *TEI Guidelines*
pour avoir une idée de toutes les choses que vous pouvez faire.

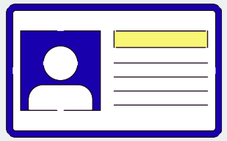
Si vous ne savez pas quel élément utiliser,
cherchez le dans les fiches via un moteur de recherches !

Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML
    a. Exercice
3. Utiliser les Guidelines de la TEI
4. Encodage : Niveau 1





NIVEAU 1 - CRÉER LE <TEIHEADER>



NIVEAU 2 - STRUCTURE PHYSIQUE



NIVEAU 3 - STRUCTURE DU CONTENU



NIVEAU 4 - CONTENU (ENTITÉS ET MÉTADONNÉES)



NIVEAU 5 - DETAILLER LE CONTENU PAR DES ATTRIBUTS

4. Encodage : Niveau 1 - Créer le <teiHeader>

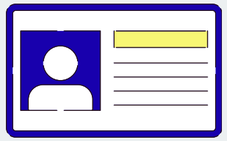
C’est quoi? C’est comme la couverture d’un livre qui contient le péritexte éditorial (auteurs, publication…)
Le <teiHeader> donne des informations sur la provenance, propriété intellectuelle, et plus encore.

Avec l’espace de noms on a vu :

- C’est la valeur de **@xmlns** qui indique à l'ordinateur la version XML utilisée.
- On utilise un attribut et une valeur ciblant l'espace de noms dans notre élément racine
  - <TEI xmlns="http://www.tei-c.org/ns/1.0">

ET <TEI> possède deux éléments obligatoires :

- Le **<teiHeader>** qui fournit les informations descriptives et déclaratives à propos du document qu’on est en train d’encoder. Il constitue une sorte de page de garde électronique, conforme à la norme imposée par la TEI.
- Le **<text>** : qu’on a déjà vu, qui contient le contenu de nos documents, nos données.



NIVEAU 1 - CRÉER LE <TEIHEADER>

C'est quoi? C'est comme la couverture d'un livre qui contient le péritexte éditorial (auteurs, publication...)

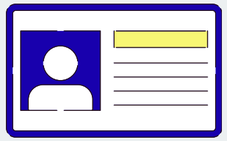
L'élément **<teiHeader>** possède un élément enfant (child) obligatoire :

<fileDesc> (la description du fichier)

1. qui contient une description bibliographique complète du fichier électronique.
2. qui doit contenir les trois éléments enfants suivants : **<titleStmt>**, **<publicationStmt>** et **<sourceDesc>**.
3. Les autres éléments sont facultatifs.

L'élément <teiHeader> possède plusieurs éléments enfants facultatifs :

1. **<encodingDesc>** (la description de l'encodage) qui documente la relation entre un texte électronique et ses sources.
2. **<profileDesc>** (description du profil du texte) qui donne une description détaillée des aspects non bibliographiques du texte, notamment les langues et sous-langues utilisées, le contexte de sa production, les participants et leur environnement.
3. **<revisionDesc>** (description de la révision) qui résume l'historique des révisions du fichier.



NIVEAU 1 - CRÉER LE <TEIHEADER>

ÉLÉMENTS OBLIGATOIRES

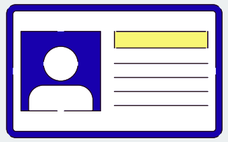
C'est quoi? C'est comme la couverture d'un livre qui contient le péritexte éditorial (auteurs, publication...)

<teiHeader>: Conteneur principal pour le péritexte éditorial

<titleStmt>: Contient les informations de l'édition numérique en cours de description, et non de l'œuvre originale
<title> : Titre du document, ex: "The Time Machine, la version numérique!"
<author> : Vos noms et prénoms à vous (puisque vous êtes les créateurs de l'édition numérique)
<editor> : Éditeur, si applicable (mais ici, vous n'en avez pas)

<publicationStmt>: Détails de publication numérique
<publisher> : Editeur (ex: disons que Atelier TEI est notre éditeur dans notre cas)
<pubPlace> : Lieu de publication
<date> : Date de publication (date de création du document, ou de publication finale)

<sourceDesc>: Description de la source de l'oeuvre originale
<bibl> : Référence bibliographique complète
<title>: Titre de l'oeuvre originale
<author>: Auteur ou artiste
<publisher>: Editeur
<date when="2019-04-12"/>:Date de publication
<biblScope unit="pages">29-40**</biblScope>**: Envergure physique couverte par l'édition numérique



NIVEAU 1 - CRÉER LE <TEIHEADER>

PRATIQUE !

C'est quoi? C'est comme la couverture d'un livre qui contient le péritexte éditorial (auteurs, publication...)

Exercice pratique de niveau 1 :

- Créez maintenant le <teiHeader> de votre document TEI
- Suivez les étapes de la fiche précédente en incluant les éléments obligatoires
- Référez-vous à la vidéo de démonstration !

Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML
    a. Exercice
3. Utiliser les Guidelines de la TEI
4. Encodage : Niveau 1
5. Encodage : Niveau 2





NIVEAU 2 – STRUCTURE PHYSIQUE

C'est quoi? La division de l'ouvrage en chapitres, les sauts de page, tous les éléments qui physiquement structurent un contenu textuel écrit ou une liste de choses.

<text> : Place principale pour le contenu
<body>: Elément obligatoire pour faire comprendre à l'ordinateur qu'il lit le corps du contenu désormais

** ** : Divisions textuelles ou artistiques auxquelles on peut donner des attributs **@types**
@type : Type de division (ex: indiquer si la division est un "chapitre", une "section", etc.)

<head> : Titre du chapitre ou de l'œuvre suivant la présentation du document ou de la liste
<p> : Paragraphes de prose ou descriptions écrites
<lg> : Groupes de lignes, surtout utilisé en poésie pour les strophes
<l> : Lignes individuelles, surtout utilisé en poésie pour les vers

<pb n="20"/> : Sauts de page auxquelles on doit donner un numéro, celui de la page, dans un attribut **@n**
<lb> : Sauts de ligne



NIVEAU 2 – STRUCTURE PHYSIQUE

C’est quoi? La division de l’ouvrage en chapitres, les sauts de page, tous les éléments qui physiquement structurent un contenu textuel écrit ou une liste de choses.

Exercice pratique de niveau 2 :

- Encodez la structure **physique** du <body> de votre document !
- Utilisez la fiche précédente pour trouver des éléments pertinents à votre document
  - Ou bien, faites des recherches sur les TEI Guidelines pour trouver les éléments qui correspondent le mieux aux besoin de vos structure physique.
- Référez-vous à la vidéo de démonstration !

Atelier III

31 mars 2026

1. Rappels importants
2. Naviguer sur l’arbre XML
    a. Exercice
3. Utiliser les Guidelines de la TEI
4. Encodage : Niveau 1
5. Encodage : Niveau 2
6. Encodage : Niveau 3





NIVEAU 3 – STRUCTURE DU CONTENU

C'est quoi? Indiquer si le contenu est un dialogue, de la prose, une description, une citation, une référence, etc.

<q>: Citations
<said>: Dialogues, paroles prononcées par les personnages, les deux fonctionnent avec **@who**
@who : Indiquer l'orateur, le personnage qui parle, etc.

<sp> : Répliques pour les dialogues ou le théâtre
<speaker> : Nom de la personne qui s'exprime à voix haute
<p> : Contenu textuel de la réplique prononcée

<note> : Notes explicatives ou annotations qui viennent avec un attribut @type
@type : Type de note (ex: "footnote", "historique")

<figure> : Conteneur pour les images ou les œuvres d'art
<graphic> : Sert à intégrer l'image et fonctionne avec les attributs suivants :
@url : Lien ou chemin vers le fichier image
@width et **@height** : Dimensions de l'image (facultatif)
<figDesc> : Description de l'image ou de l'œuvre (ex: son sujet, ce qui est représenté = iconographie, style, etc.)



NIVEAU 3 - STRUCTURE DU CONTENU

C'est quoi? Indiquer si le contenu est un dialogue, de la prose, une description, une citation, une référence, etc.

Exercice pratique de niveau 3 :

- Encodez la structure du **contenu** de votre document !
- Utilisez la fiche précédente pour trouver des éléments pertinents à votre document
  - Ou bien, faites des recherches sur les TEI Guidelines pour trouver les éléments qui correspondent le mieux aux besoin de vos structure physique.
- Référez-vous à la vidéo de démonstration !

Introduction au langage XML pour l'analyse de texte

une série d'ateliers

Les mardis de 10 h 30 à 12 h (FR)
Les mardis de 13 h 30 à 15 h (EN)

Avec Alice Defours et
Constance Crompton

