

Introduction to TEI

Matthias GILLE LEVENSON

Université Aix-Marseille, École Normale Supérieure de Lyon, France

matthias [dot] gille [-] levenson [at] ens-lyon [dot] fr

June 3rd, 2025



- 1 Theoric introduction
- 2 TEI Fundamentals
- 3 Conclusion

- 1 Theoric introduction
 - What is the TEI?
 - A little bit of history
 - Principles
 - The TEI, what for ?
 - The XML format

What is the TEI?

- A community
- A standard
- A way of seeing/modelling “the” text

1 Theoric introduction

- What is the TEI?
- A little bit of history
- Principles
- The TEI, what for ?
- The XML format

- Born in 1987
- 5 versions (actual: P5)
- Originally: SGML; switch to XML in 2007

1 Theoric introduction

- What is the TEI?
- A little bit of history
- Principles
- The TEI, what for ?
- The XML format

■ Separate the appearance and the “essence” of textual objects

del mal del otro o del buen zelo³. Si del mal, cuidando que non devía el omne sufrir tal mal, esta es misericordia, ca misericordia non es otra cosa sinon tristeza del mal que omne sufre sin merescimiento, segund que dize el philósofo en el IIº de la *Retórica*; mas si toma tristeza del bien, esto es en dos maneras, ca o le pesa del bien que otri ha, maguer que lo él meresca de aver, e esta es envidia, ca envidia non es otra cosa sinon // [Fol. 176r] dolor o tristeza del bien que otro ha⁴; [mas si le pesa del bien que otro ha] porque lo non meresçe aver, (e) así es némesis o

³ del buen zelo] del bien *Glosa*.

⁴ En este punto ha producido un salto de texto de igual a igual respecto a la *Glosa*, donde se lee seguidamente el texto que transcribo entre corchetes, que es necesario para que el sentido del pasaje no quede truncado.

⁵ [a los mayores]] *om. Castigos* respecto a la *Glosa*.

Figure: Fragment of and edition of the *Castigos de Sancho IV*

■ Separate the appearance and the “essence” of textual objects

```

<p>del mal del otro o
  <app>
    <lem>del buen zelo</lem>
    <rdg wit="#Glosa">del bien</rdg>
  </app>
  . Si del mal, cuidando que non devía el omne sufrir tal mal, esta es misericordia, ca
  misericordia non es otra cosa sinon tristeza del mal que omne sufre sin merescimiento,
  segund que dize el philósofo en el llo de la Retórica; mas si toma tristeza del bien,
  esto es en dos maneras, ca o le pesa del bien que otri ha, maguer que lo él meresca de
  aver, e esta es envidia, ca envidia non es otra cosa sinon <pb n="176r"/> dolor o
  tristeza del bien que otro ha;
  <choice>
    <sic>mas si le pesa del bien que otro ha</sic>
    <corr/>
  </choice>
  <note>En este punto ha producido un salto de texto de igual a igual respecto a la Glosa,
  donde se lee seguidamente el texto que transcribo, que es necesario
  para que el sentido del pasaje no quede truncado.</note> porque lo non meresçe aver,
  <supplied>e</supplied> así es némesis o desdén</p>

```

Figure: Its possible representation in XML-TEI

- 1 Theoric introduction
 - What is the TEI?
 - A little bit of history
 - Principles
 - The TEI, what for ?
 - The XML format

The TEI, what for ?

- Describing a text using the experience of a large community
- Producing semantic data that can be read by the human and by the computer
- Easing documents sharing and reusability

1 Theoric introduction

- What is the TEI?
- A little bit of history
- Principles
- The TEI, what for ?
- The XML format

Conformance and validity

- XML stands for **eXtensible Markup Language**.
- It is a format that allows to describe textual elements

Conformance and validity

- Two important concepts
- A document **must** be XML conformant, that is, respect the rules of the XML format
- A document **may** be validated against a schema, that is a document that verifies some rules are respected
- Some examples of specifications: TEI, EAD, DublinCore, AltoXML, PageXML, RDF, etc...

Conformance

■ Elements, attributes, values

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xml:id="drp" xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <author cert="low">Matthias Gille Levenson</author>
      </titleStmt>
      <publicationStmt>
        <publisher>Unpublished</publisher>
        <availability>
          <licence>Creative Commons CC BY-NC-SA 4.0 FR</licence>
        </availability>
      </publicationStmt>
    </fileDesc>
  </teiHeader>
</TEI>
```

One node to contain them all

- One and only one top element that contains everything else

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xml:id="drp" xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <author cert="low">Matthias Gille Levenson</author>
      </titleStmt>
      <publicationStmt>
        <publisher>Unpublished</publisher>
        <availability>
          <licence>Creative Commons CC BY-NC-SA 4.0 FR</licence>
        </availability>
      </publicationStmt>
    </fileDesc>
  </teiHeader>
</TEI>
```


Conformance

- No overlapping elements

```
<?xml version="1.0" encoding="UTF-8"?>  
<TEXT xml:id="drp" xmlns="http://www.tei-c.org/ns/1.0">  
  <NODE_A>Hello <NODE_B>World</NODE_A></NODE_B>  
</TEXT>
```

Conformance

- Some elements can contain other elements, but elements can also be empty

```
<?xml version="1.0" encoding="UTF-8"?>
<div type="chapitre" n="1">
  <lb/>ca . i. quomodo diuiduntur potentiae
  <lb/>anime et in quibus potenciis
  <lb/>habent esse
  <lb/>uirtutes
</div>
```

Conformance

- Attribute values must be inside quotes

```
<?xml version="1.0" encoding="UTF-8"?>
<TEXT>
  <NODE_A type=sentence>Hello
    <NODE_B>World</NODE_B>
  </NODE_A>
</TEXT>
```

Conformant or not ?

```

1 <sentence>Longtemps, je me suis couché de bonne heure.</sentence>
2
3 <sentence>Longtemps, je me suis couché de bonne heure.<sentence>
4
5 <sentence type=incipit>Longtemps, je me suis couché<linebreak></linebreak> de bonne heure.</sentence>
6
7 <sentence type=incipit>Longtemps, je me suis couché<linebreak/> de bonne heure.</sentence>
8
9 <sentence>Longtemps, je me suis couché de bonne heure.</sent>
10
11 <sentence type="incipit">Longtemps, je me suis couché de bonne heure.</sentence>
12
13 <paragraph>
14     <sentence type="incipit">Longtemps, je me suis couché <striketrough>de bonne heure.</sentence>
15     <sentence>Parfois, à peine ma bougie éteinte</striketrough>,
16         mes yeux se fermaient si vite que je n'avais pas le temps de me dire</sentence>
17 </paragraph>
18
19 <paragraph>
20     <sentence type="incipit">Longtemps, je me suis couché <striketrough>de bonne heure.</striketrough></sentence>
21     <sentence><striketrough>Parfois, à peine ma bougie éteinte</striketrough>,
22         mes yeux se fermaient si vite que je n'avais pas le temps de me dire</sentence>
23 </paragraph>
24

```

Schemas

- A schema is a document that is used to control the quality of some encoding.

Schemas

```

TEI text body div div div lb
1803 <body>
1804 <div type="livre" n="1">
1805 <div type="partie" n="2">
1806 <div type="chapitre" n="1"><pb n="" facs="..input_files/16369405.jpg"/><fw
1807 type="titre_courant"/><lb break="?" rend="rubric"
1808 xml:id="elem_eSc_line_99cc2a93"/>Capitulum . i. Iquno diuiduntur potencie<lb
1809 break="yes" rend="rubric"
1810 xml:id="elem_eSc_line_7f955f55"/>anime et in quibus potenciis habet esse uirtutes.<lb
1811 break="yes"
1812 xml:id="elem_eSc_line_6ea8e267"/>ostquam auxiliante deo compleui<lb break="no"
1813 xml:id="elem_eSc_line_df9d9bc9"/>mus primampartem huius primi libri inquo<lb
1814 break="yes"
1815 xml:id="elem_eSc_line_b36db744"/>agitur de regimine sui ostendentes<lb break="no"
1816 xml:id="elem_eSc_line_2c352677"/>i quo reges et pñcipes suam felici [] poner<lb
1817 break="no"
1818 xml:id="elem_eSc_line_54b88267"/>e d[] quia non decet eos suum finem pone in din<lb
1819 break="no"
1820 xml:id="elem_eSc_line_a6468504"/>ciis. nec in ciuili potentia. nec in aliquibus
1821 talibus sed omni<lb break="no"
1822 xml:id="elem_eSc_line_5efa6adc"/>bus hiis ut supra plenius perba debet uti tanquam<lb
1823 break="yes"
1824 xml:id="elem_eSc_line_33a63380"/>organis ad felicitatem sed ponedz in a<lb
1825 break="no"
1826 xml:id="elem_eSc_line_c991f46c"/>ctu pruden [] prout talis actus est inpatu[]<lb
1827 break="yes"
1828 xml:id="elem_eSc_line_388d53ce"/>a cari [] Nam tunc reges habet felicitat<lb
1829 break="no"
1830 xml:id="elem_eSc_line_4b955e3e"/>tem suo statui debitas et condignam quando<lb
1831 break="yes"
1832 xml:id="elem_eSc_line_c422ed80"/>instigante dei dilectione secundum prudenciam<lb
1833 break="yes"

```

❗ text not allowed here; expected the element end-tag or element "ab", "addSpan", "alt", "altGrp", "anchor", "annotationBlock", "app", "argument", "bibl", "biblFull", "biblStruct", "byline", "camera", "caption", "castList",

Figure: This fragment is well-formed, but it is not TEI compliant / valid. A paragraph p or an anonymous block ab should wrap the lines.

Schemas

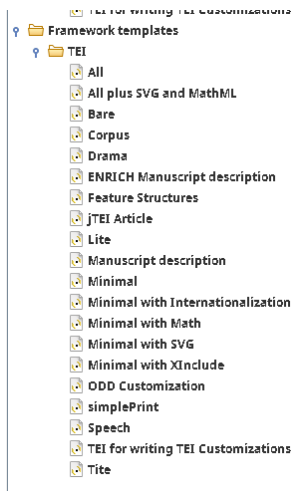


Figure: There are multiple TEI schemas available, adapted to the user's usecases

ODD's


- ODD stands for “One Document Does it All”
- An ODD is a TEI document that is used to create documentation and schema, with a transformation script maintained by the TEI community

- 1 Theoric introduction
- 2 TEI Fundamentals**
- 3 Conclusion

The TEI guidelines

- <https://tei-c.org/guidelines/>
- The TEI guidelines is the document you will consult everyday starting now. You can print it and use it as a bedside reading.
- The guidelines are made of two main parts:
 - a description of good editing practices, in natural language
 - the individual description of each element (possible attributes, elements inside, elements containing the current element, etc.)

The TEI guidelines


Guidelines ▾ Activities ▾ Tools ▾ Membership ▾ Support ▾ About ▾ News

TEI: Guidelines for Electronic Text Encoding and Interchange
P5 Version 4.9.0. Last updated on 24th January 2025, revision f73186978

Table of contents
12.1 Digital Facsimiles
12.2 Combining Transcription with Facsimile
12.3 Scope of Transcriptions
12.4 Aspects of Layout
12.5 Transcription and Ruby
12.6 Headers, Footers, and Similar Matter
12.7 Identifying Changes and Revisions
12.8 Other Primary Source Features not Covered in these Guidelines
12.9 Module for Transcription of Primary Sources

12.1 Manuscript Description
13 Critical Apparatus
Home

12 Representation of Primary Sources

This chapter describes elements that may be used to represent primary source materials, such as manuscripts, printed books, ephemera, or other textual documents. Some of these specialized elements, particularly at phrase-level, add to the other elements available within [text](#) to deal with textual phenomena more specific to primary source transcription. Other structural and block-level elements described here can be used to represent primary source materials by prioritizing the encoding of their spatial features over their logical textual structure (that is, the elements described in chapter 4 [Default Text Structure](#)). These elements, [facsimile](#), [sourceDoc](#), and their children, may be used in parallel and in combination with an encoding of logical text structures with [text](#), or as standalone representations. The element [sourceDoc](#) in particular provides a way of combining facsimile and transcriptions by embedding transcribed text. This approach focuses on physical and textual features that can be primarily described spatially, such as the sequence of pages in a manuscript, or the layout of a printed page. This is not meant to be the only way of transcribing primary sources in TEI, or even a preferred way; which approach is more appropriate will depend on the specific needs of your project.

Although this chapter discusses manuscript materials more frequently than other forms of written text, most of the recommendations presented are equally applicable to facsimiles of a wide variety of media, including printed matter, monumental inscriptions, and art. Each medium has its own vocabulary of agents. In the following examples, terms such as 'scribe', 'author', 'editor', 'annotator' or 'corrector' may be re-interpreted in terms more appropriate to the medium being transcribed. In printed material, for example, the 'compositor' plays a role analogous to the 'scribe', while in an authorial manuscript, the 'author' and the 'scribe' are the same person.

This module may be used in conjunction with other modules. These recommendations are not intended to meet every transcriptional circumstance likely to be faced by any scholar. They are intended rather as a base to enable encoding of the most common phenomena found in the course of scholarly transcription of primary source materials. These guidelines do not address the encoding of physical description of textual witnesses: the materials of the carrier, the medium of the inscribing implement, the organisation of the carrier materials themselves (such as quiring, collation), authorial instructions or scribal markup, etc., except insofar as these are involved in the broader question of manuscript description, as addressed by the [msdescription](#) module described in chapter 11 [Manuscript Description](#).

This chapter begins by describing elements for handling digitally-encoded images of primary source materials for the purpose of creating digital facsimiles using the [facsimile](#) element (12.1 [Digital Facsimiles](#)).

The next section (12.2 [Combining Transcription with Facsimile](#)) describes two ways of combining a facsimile images with a transcription; either by referencing a parallel transcription in [text](#), or by providing an 'embedded' transcription that prioritizes the encoding of a resource's spatial features via the [sourceDoc](#) element and a number of transcriptional elements.

Section 12.3 [Scope of Transcriptions](#) documents elements that support scholars in recording information about specific features of the text written on its physical carrier, such as 12.3.1 [Altered, Corrected, and Erroneous Texts](#) and 12.3.2 [Hands and Responsibility](#).

Section 12.4 [Aspects of Layout](#) describes how complex page layouts may be represented.

Section 12.6 [Headers, Footers, and Similar Matter](#) introduces the element [fw](#) (forme work) for encoding material repeated from page to page that falls outside the stream of the text.

Section 12.7 [Identifying Changes and Revisions](#) describes how to document changes made during the production or revision of a primary source.

The chapter concludes with a technical overview of the structure and organization of the module described here. Some elements from other chapters are recontextualized for situations involving the transcription of primary source materials, whether within [text](#) or [sourceDoc](#). Therefore, this overview should be read in conjunction with chapters 3 [Elements Available in All TEI Documents](#) and 9 [Characters, Glyphs, and Writing Modes](#).

12.1 Digital Facsimiles

12.2 Combining Transcription with Facsimile
Home



12.1 Digital Facsimiles

A common approach in the TEI to representing pre-existing sources involves transcribing or otherwise converting sources into character form before marking them up. However, it is also a common practice to make a different form of 'digital text' that is instead composed of digital images of the original source, typically one per page, or other written surface. We call such a resource a *digital facsimile*. A digital facsimile may, in the simplest case, just consist of a collection of images, with some metadata to identify them and the source materials portrayed. It may sometimes contain a variety of images of the same source pages, perhaps of different resolutions, or of different kinds. Such a collection may form part of any kind of document, for example a commentary of a codicological or paleographic nature, where there is a need to align

12.2 Combining Transcription with Facsimile

Home

Figure: The natural language part of the TEI

The TEI guidelines


[Guidelines](#) ▾
 [Activities](#) ▾
 [Tools](#) ▾
 [Membership](#) ▾
 [Support](#) ▾
 [About](#) ▾
 [News](#)

▾

TEI: Recommandations pour l'encodage et l'échange de textes électroniques

P5 Version 4.9.0. Last updated on 24th January 2025, revision f73186978

<lb>

Accueil
C. Éléments

<lb> (début de ligne) marque le début d'une nouvelle ligne (typographique) dans une édition ou dans une version d'un texte. [3.11.3 Milestone Elements 7.2.5 Speech Contents]	
Module	core — Elements Available in All TEI Documents
Attributs	<ul style="list-style-type: none"> att.global: @xml:id, @n, @xml:lang, @xml:base, @xml:space
Membre du	model.milestoneLike
Contenu dans	analysis: cl m p hr s span w transcr: damage fw line metamark mod restore retrace secl sourceDoc subst supplied surface surfaceGrp surplus zone verse: metSym rhyme
Peut contenir	Élément vide
Note	<p>Par convention, l'élément lb apparaît à l'endroit du texte où commence une nouvelle ligne. L'attribut @n, s'il est utilisé, donne un nombre ou une autre valeur associée au texte entre ce point et l'élément suivant lb, spécifiquement le numéro de la ligne dans la page, ou une autre unité de mesure appropriée. Cet élément est prévu pour être employé pour marquer un saut de ligne sur un manuscrit ou sur une page imprimée, à l'endroit où il se survient; on n'utilisera pas de balisage structurel comme une succession de vers (pour lequel l'élément l est disponible) sauf dans le cas où des blocs structurés ne peuvent pas être marqués autrement.</p> <p>L'attribut @type sera employé pour caractériser toute espèce de caractéristiques du saut de ligne, sauf la coupure des mots (indique par l'attribut @break) ou la source concernée.</p>

Figure: The description of the lb element (simplified to fit in the slide)

2 TEI Fundamentals

■ Main components

■ Structure

■ Document layout description and material description of sources

■ Document layout description and material description of sources

■ Exercise 1. Encoding a poem.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Title</title>
      </titleStmt>
      <publicationStmt>
        <p>Publication information</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Some text here.</p>
      <figure>
        <graphic url="http://www.tei-c.org/logos/TEI-glow.png"/>
      </figure>
    </body>
  </text>
</TEI>
```

Figure: The minimal TEI document

- Two main components: data and metadata

The `teiHeader` element

- The `teiHeader` contains the metadata and all the information about the sources you are describing.
- Four main components: `fileDesc`, `encodingDesc`, `profileDesc`, `revisionDesc`.

The fileDesc element

- The fileDesc contains the bibliographic information about the source.

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>
        <!-- title of the resource -->
      </title>
    </titleStmt>
    <publicationStmt>
      <p>
        <!-- Information about distribution of the resource -->
      </p>
    </publicationStmt>
    <sourceDesc>
      <p>
        <!-- Information about source from which the resource derives -->
      </p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Figure: Example of the fileDesc taken from the TEI Guidelines

The encodingDesc element

- The encodingDesc is used to describe the principles that the editor has been following to produce the TEI document.

```
<encodingDesc>
  <projectDesc>
    <p>Originally prepared for use in the production of a series of old-spelling
      concordances in 1968, this text was extensively checked and revised for use during
      the
      editing of the new Oxford Shakespeare (Wells and Taylor, 1989).</p>
  </projectDesc>
  <editorialDecl>
    <correction>
      <p>Turned letters are silently corrected.</p>
    </correction>
    <normalization>
      <p>Original spelling and typography is retained, except that long s and ligatured
        forms are not encoded.</p>
    </normalization>
  </editorialDecl>
  <refsDecl xml:id="ASLREF">
    <cRefPattern matchPattern="(\S+) ([^.]*)\.(.*)"
      replacementPattern="#xpath(//div1[@n='S1']/div2[@n='S2']/lb[@n='S3'])">
      <p>A reference is created by assembling the following, in the reverse order as that
        listed here: <list>
          <item>the <att>n</att> value of the preceding <gi>lb</gi>
          </item>
          <item>a period</item>
          <item>the <att>n</att> value of the ancestor <gi>div2</gi>
          </item>
          <item>a space</item>
          <item>the <att>n</att> value of the parent <gi>div1</gi>
          </item>
        </list>
      </p>
    </cRefPattern>
  </refsDecl>
</encodingDesc>
```

Figure: Example of the encodingDesc taken from the TEI Guidelines

The text

- The `text` element contains the text *per se*. It contains three main elements: `front`, `body`, `back`

Facsimiles

- Facsimile information is stored in a specific element after the `teiHeader`, in `facsimile` element. It is an element that's being used everyday more and more due to the apparition of efficient HTR algorithms.
- This element tends to be created automatically, as it can contain lots and lots of subelements (page and line

Facsimiles

```
<facsimile>
  <surface ulx="0" uly="0" lrx="200" lry="300">
    <graphic url="Bovelles-49r.png"/>
    <zone ulx="25" uly="25" lrx="180" lry="60">
      <!-- contains the title -->
    </zone>
    <zone ulx="28" uly="75" lrx="175" lry="178"/>
      <!-- contains the paragraph in italics -->
    <zone ulx="105" uly="76" lrx="175"
      lry="160"/>
      <!-- contains the figure -->
    <zone ulx="45" uly="125" lrx="60" lry="130"/>
      <!-- contains the word "pendans" -->
    </surface>
  </facsimile>
```

Figure: The description of the elements on the page with their coordinates

2 TEI Fundamentals

- Main components

- Structure

- Document layout description and material description of sources
- Document layout description and material description of sources
- Exercise 1. Encoding a poem.

Some basic structuring elements

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

- `div`: for any division of a text. They can be nested, and attributes like `@type` or `@n` are used to specify the type and level of structure
- `head`: for encoding headings
- `p`: for encoding paragraphs
- `ab` (for anonymous block): for encoding any sub-div block
- `lg` (line group)
- `l` (verse)

2 TEI Fundamentals

- Main components
- Structure
- Document layout description and material description of sources
- Document layout description and material description of sources
- Exercise 1. Encoding a poem.

Some basic structuring elements

`https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html`

- pb: page beginning
- cb: columns beginning
- lb: line beginning
- fw (forme work): for encoding headers, footers, page number, catchwords, etc

2 TEI Fundamentals

- Main components
- Structure
- Document layout description and material description of sources
- Document layout description and material description of sources
- Exercise 1. Encoding a poem.

Some basic structuring elements

```

<div type="livre" n="1">
  <div type="partie" n="2">
    <div type="chapitre" n="1"><ab><pb n="" facs="../input_files/16369405.jpg"/><fw
      type="titre_courant"/><lb break="?" rend="rubric"
      xml:id="elem_eSc_line_99cc2a93"/>Capitulum . i. Iquno diuiduntur potencie<lb
      break="yes" rend="rubric"
      xml:id="elem_eSc_line_7f955f55"/>anime et in quibus potenciis habet esse uirtutes. <lb
      break="yes"
      xml:id="elem_eSc_line_6ea8e267"/>ostquam auxiliante deo compleui<lb break="no"
      xml:id="elem_eSc_line_df9d9bc9"/>mus primampartem huius primi libri inquo<lb
      break="yes"
      xml:id="elem_eSc_line_b36db744"/>agitur de regimine sui ostendentes<lb break="no"
      xml:id="elem_eSc_line_2c352677"/>i quo reges et p̃ncipes suam felici ☐ ☐ poner<lb
      break="no"
      xml:id="elem_eSc_line_54b88267"/>e d̃ ☐ quia non decet eos suum finem pone in din<lb
      break="no"
      xml:id="elem_eSc_line_a6468504"/>ciis. nec in ciuili potentia. nec in aliquibus
      talibus sed omni<lb break="no"
      xml:id="elem_eSc_line_5efa6adc"/>bus hiis ut supra plenius perba debet uti tanquam<lb

```

2 TEI Fundamentals

- Main components
- Structure
- Document layout description and material description of sources
- Document layout description and material description of sources
- Exercise 1. Encoding a poem.

Shakespeare's Sonnet 18

- Encode the Sonnet 18, available in the Git Repo
- The “verse” section of the guidelines will help you here:
<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html>

- 1 Theoric introduction
- 2 TEI Fundamentals
- 3 Conclusion**

- TEI-based editions without publication of the XML sources is quite unuseful
- Please publish your data alongside its documentation !

3 Conclusion

- What's next ? Manipulating XML trees

XPath

- XPath is the base language to navigate trees

```
<xsl:template  
  match="tei:TEI[@type = 'transcription'][not(@subtype = 'version_a')][not(descendant::tei:text[@xml:lang = 'la'])]">
```

Figure: A (kind of) simple XPath query, inside quotes

XSLT

- XSLT is a transformation language that is built on the same logic as XML: nesting
- It is useful for creating complex (web-based or \LaTeX) editions

```
<xsl:template
  match="text()"
  mode="secondePasse">
  <xsl:for-each select="tokenize(., '\s+')">
    <xsl:analyze-string select="." regex="([()::;?!\.])">
      <xsl:matching-substring>
        <xsl:element name="pc" namespace="http://www.tei-c.org/ns/1.0">
          <xsl:value-of select="regex-group(1)"/>
        </xsl:element>
      </xsl:matching-substring>
      <xsl:non-matching-substring>
        <xsl:element name="w" namespace="http://www.tei-c.org/ns/1.0">
          <xsl:value-of select="."/>
        </xsl:element>
      </xsl:non-matching-substring>
    </xsl:analyze-string>
  </xsl:for-each>
</xsl:template>
```

Figure: A rule to tokenize a text into words and punctuation with XSLT

XQuery

- XQuery is a query language that is used to build XML databases and create dynamic editions.

```
declare namespace tei = "http://www.tei-c.org/ns/1.0";  
for $stones in collection('/db/pc')//tei:TEI[@n < 11]  
let $stoneID := $stones/@xml:id  
let $recordTitle := $stones//tei:titleStmt/tei:title  
return  
  <ul>  
    <li>Stone ID: {data($stoneID)} </li>  
    <li>Record Title: {data($recordTitle)} </li>  
  </ul>
```

Figure: Example of query on a collection of tombstones inscriptions encoded in XML-TEI. Taken from James Cummings' workshop at TEI Conference 2006.

Python

- Python is usefull to plug external tools for text processing. It is really helpfull for NLP tasks (annotation, segmentation, etc)
- It can be more performant for treating large corpora than XSLT/XQuery
- It follows a linear logic and therefore is not adapted to in-depth transformations of XML sources (not suited for editing for instance)

```
with open(xml_file, "r") as input_file:
    f = etree.parse(input_file)
    line_breaks = f.xpath("//tei:lb[not(parent::tei:fw)]", namespaces=namespace_declaration)
    text_lines = [utils.clean_and_normalize_encoding(line.tail) for line in line_breaks]
    predictions = []
    steps = len(text_lines) // batch_size
    for n in range(steps):
        batch = text_lines[n * batch_size: (n * batch_size) + batch_size]
        predicted_batch = tagger.tag_and_detect_lb(batch)
        predictions.extend(predicted_batch)
    predictions.extend(tagger.tag_and_detect_lb(text_lines[(n + 1) * batch_size:]))
```

Figure: Some code that extracts each line of a transcription to detect hyphenated lines



Baillot, Anne and Julie Giovacchini. “TEI Models for the Publication of Social Sciences and Humanities Journals: Opportunities, Challenges, and First Steps Toward a Standardized Workflow”. In: *Journal of the Text Encoding Initiative* 14 (2021).



Burnard, Lou. *Qu'est-ce que la Text Encoding Initiative ?* OpenEdition Press, 2015. ISBN: 978-2-8218-5580-9.



— . “What Is TEI Conformance, and Why Should You Care?” In: *Journal of the Text Encoding Initiative* 12 (2019). ISSN: 2162-5603.



Camps, Jean-Baptiste. *TEI, LaTeX et Les Éditions Critiques Sur Papier : I. Les Différents Packages*. Sacré Gr@@l. 2011. URL: <http://graal.hypotheses.org/484> (visited on 12/07/2017).



Eide, Øyvind. “Ontologies, Data Modeling, and TEI”. In: *Journal of the Text Encoding Initiative* (Issue 8 Dec. 28, 2014). ISSN: 2162-5603. DOI: 10.4000/jtei.1191.



Fradejas Rueda, José Manuel. “La Codificación TEI de Las Ediciones de 1491 y 1555 de Las Siete Partidas”. In: *Las "Siete Partidas" Del Rey Sabio: Una Aproximación Desde La Filología Digital y Material*. Iberoamericana, 2021, pp. 253–265.



Scheithauer, Hugo, Alix Chagué, and Laurent Romary. “From eScriptorium to TEI Publisher”. In: *Brace your digital scholarly edition!* (2021).



Torsten Roeder, Leopoldina. “Juxta Web Service, LERA, and Variance Viewer. Web Based Collation Tools for TEI”. In: *RIDE* 11 (2020). DOI: 10.18716/ride.a.11.5.