

Modelling humanities data with TEI-XML

SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Dr Katarzyna Anna Kapitan
29 October 2025

<msDesc>: How it works

```
<msDesc>
    <msIdentifier> </msIdentifier>
    <msContents>
        <msItem>
            <author> </author>
            <title> </title>
        </msItem>
    </msContents>
    <physDesc>
        <objectDesc>
            <supportDesc> </supportDesc>
            <layoutDesc> </layoutDesc>
        </objectDesc>
        <handDesc> </handDesc>
        <bindingDesc> </bindingDesc>
    </physDesc>
    <history> </history>
    <additional> </additional>
</msDesc>
```

<physDesc>

- ▶ <**physDesc**> describes aspects of the form, support, extent, and quire structure of the manuscript object and of the way in which the text is laid out on the page
- ▶ the styles of writing, decorative features, any musical notation employed and any annotations or marginalia
- ▶ discussion of its binding, seals, and any accompanying material

<physDesc>

- ▶ **<objectDesc>** contains a description of the physical components making up the object which is being described.
- ▶ **<handDesc>** contains a description of all the different hands used in a manuscript or other object.
- ▶ **<scriptDesc>** contains a description of the scripts used in a manuscript.
- ▶ **<decoDesc>** contains a description of the decoration of a manuscript.
- ▶ **<additions>** contains a description of any significant additions found within a manuscript or other object, such as marginalia or other annotations.
- ▶ **<bindingDesc>** describes the present and former bindings of a manuscript.

<objectDesc>

- ▶ <**objectDesc**> contains a description of the physical components making up the object which is being described.
 - ▶ **@form** a short project-specific name identifying the physical form of the carrier, for example as a codex, roll, fragment, partial leaf, cutting etc.
- ▶ <**supportDesc**> groups elements describing the physical support for the written part of a manuscript or other object.
 - ▶ **@material** a short project-defined name for the material composing the majority of the support.
- ▶ <**layoutDesc**> collects the set of layout descriptions applicable to a manuscript or other object.

<supportDesc>

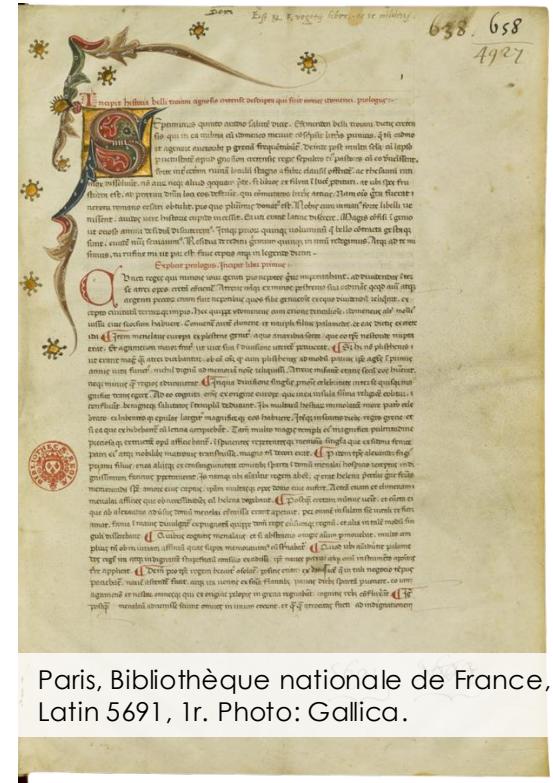
- ▶ **<support>** contains a description of the materials etc. which make up the physical support for the written part of a manuscript.
- ▶ **<extent>** describes the approximate size of a text stored on some carrier medium or of some other object, digital or non-digital, specified in any convenient units.
- ▶ **<foliation>** describes the numbering system or systems used to count the leaves or pages in a codex or similar object.
- ▶ **<collation>** contains a description of how the leaves, bifolia, or similar objects are physically arranged.
- ▶ **<condition>** contains a description of the physical condition of the manuscript or object.

Exercise 1.1: Paris, BnF, Latin 5691

File in Week4/Ex1

► Answer the following questions:

- What is the format of the manuscript? How is this indicated?
- What is the material used for support? How is this indicated?
- What is the extent of the manuscript (i.e. how many leaves are there in the manuscript)? How is this indicated?
- What are the dimensions of the manuscript? How is this indicated?
- Is the manuscript foliated? How is this indicated?
- How many gatherings there are in the manuscript? How is this indicated?



Paris, Bibliothèque nationale de France,
Latin 5691, 1r. Photo: Gallica.

<layoutDesc>

- ▶ **<layout>** describes how text is laid out on the page or surface of the object, including information about any ruling, pricking, or other evidence of page-preparation techniques.
- ▶ **@columns** specifies the number of columns per page.
- ▶ **@ruledLines** specifies the number of ruled lines per column.
- ▶ **@writtenLines** specifies the number of written lines per column.

Other elements of <physDesc>

- ▶ **<dimensions>** contains a dimensional specification.
- ▶ **<catchwords>** describes the system used to ensure correct ordering of the quires or similar making up a codex, typically by means of annotations at the foot of the page.
- ▶ **<signatures>** contains discussion of the leaf or quire signatures found within a codex or similar object.
- ▶ **<secFol>** (second folio) marks the word or words taken from a fixed point in a codex (typically the beginning of the second leaf) in order to provide a unique identifier for it.
- ▶ **<watermark>** contains a word or phrase describing a watermark or similar device.
- ▶ **<stamp>** contains a word or phrase describing a stamp or similar device.

Other elements of <physDesc>

<**handDesc**> with <**handNote**> describes a particular style or hand distinguished within a manuscript.

<**scriptDesc**> with <**scriptNote**> describes a particular script distinguished within the description of a manuscript or similar resource.

<**decoDesc**> with <**decoNote**> contains a note describing either a decorative component of a manuscript or other object, or a fairly homogenous class of such components.

<bindingDesc>

<binding> contains a description of one binding, i.e. type of covering, boards, etc. applied to a manuscript or other object.

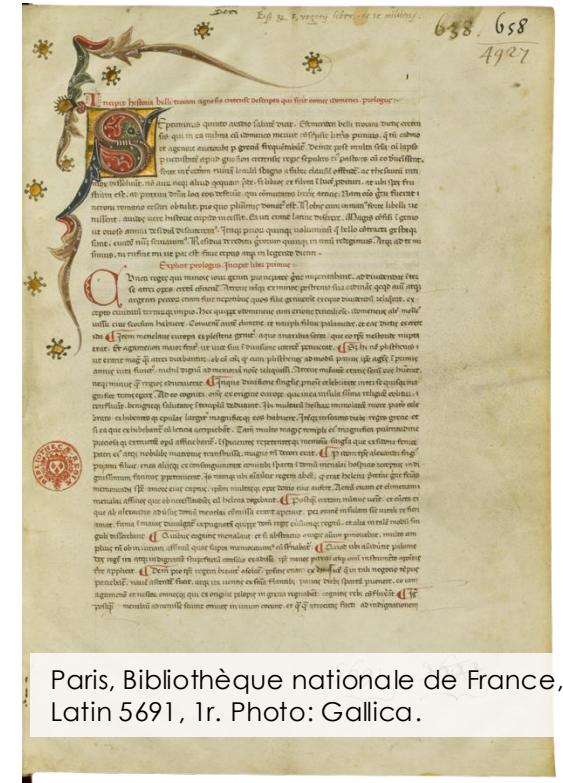
@contemporary specifies whether or not the binding is contemporary with the majority of its contents.

- ▶ The following elements can also be used:
- ▶ **<p> <condition> <decoNote>**

Exercise 1.2: Paris, BnF, Latin 5691

File in Week4/Ex1

- ▶ Answer the following questions:
 - ▶ How many columns, written lines, and rules lined are there in the manuscript?
 - ▶ What do we know about the the styles of writing and decorative features of this manuscript?
 - ▶ Who wrote this manuscript?
 - ▶ Are there any marginalia in the manuscript?
 - ▶ What is the mistake that we've made in describing the binding?



Paris, Bibliothèque nationale de France,
Latin 5691, 1r. Photo: Gallica.

<msDesc>: How it works

```
<msDesc>
    <msIdentifier> </msIdentifier>
    <msContents>
        <msItem>
            <author> </author>
            <title> </title>
        </msItem>
    </msContents>
    <physDesc>
        <objectDesc>
            <supportDesc> </supportDesc>
            <layoutDesc> </layoutDesc>
        </objectDesc>
        <handDesc> </handDesc>
        <bindingDesc> </bindingDesc>
    </physDesc>
    <history> </history>
    <additional> </additional>
</msDesc>
```

<history>

- ▶ **<origin>** contains any descriptive or other information concerning the origin of a manuscript.
- ▶ **<provenance>** contains any descriptive or other information concerning a single identifiable episode during the history of a manuscript after its creation but before its acquisition.
- ▶ **<acquisition>** contains any descriptive or other information concerning the process by which a manuscript entered the holding institution.

<origin>

- ▶ **<origDate>** contains any form of date, used to identify the date of origin for a manuscript or a manuscript part.
- ▶ **<origPlace>** contains any form of place name, used to identify the place of origin for a manuscript or a manuscript part.
- ▶ **<bibl>** contains a loosely-structured bibliographic citation of which the sub-components may or may not be explicitly tagged.

<additional>

- ▶ <**adminInfo**> contains information about the present custody and availability of the manuscript or other object, and also about the record description itself.
- ▶ <**surrogates**> contains information about any representations of the manuscript or other object being described which may exist in the holding institution or elsewhere.
- ▶ <**bibl**> contains a loosely-structured bibliographic citation of which the sub-components may or may not be explicitly tagged.
- ▶ <**listBibl**> contains a list of bibliographic citations of any kind.

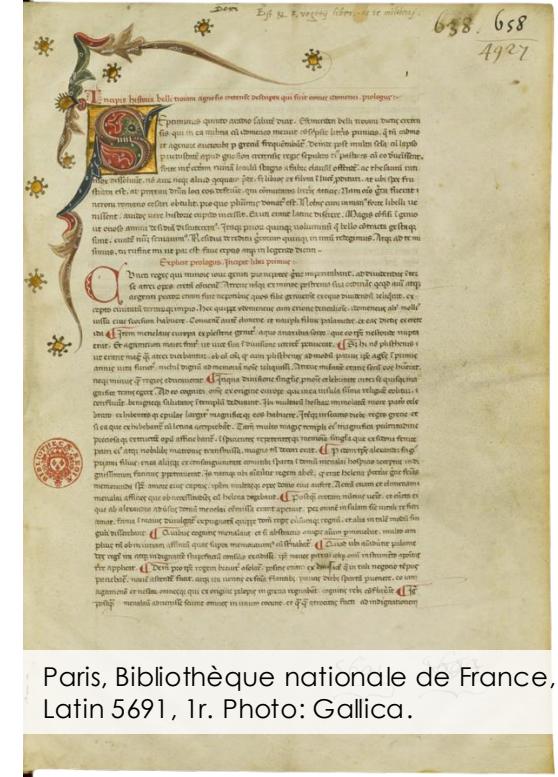
<adminInfo>

- ▶ **<recordHist>** provides information about the source and revision status of the parent manuscript or object description itself.
- ▶ **<source>** describes the original source for the information contained with a manuscript or object description.
- ▶ **<change>** documents a change or set of changes made during the production of a source document, or during the revision of an electronic file.
- ▶ **<availability>** supplies information about the availability of the manuscript, for example any restrictions on its use.
- ▶ **<custodialHist>** contains a description of a manuscript or other object's custodial history, either as running prose or as a series of dated custodial events.

Exercise 1.3: Paris, BnF, Latin 5691

File in Week4/Ex1

- ▶ Answer the following questions:
 - ▶ Who created this manuscript description? How is it indicated?
 - ▶ Who is Angelo de Sanguineis de Roma?
 - ▶ What is the mistake that we've made in describing his role?





XPath 2.0 ▾ Execute XPath on 'Current Fil



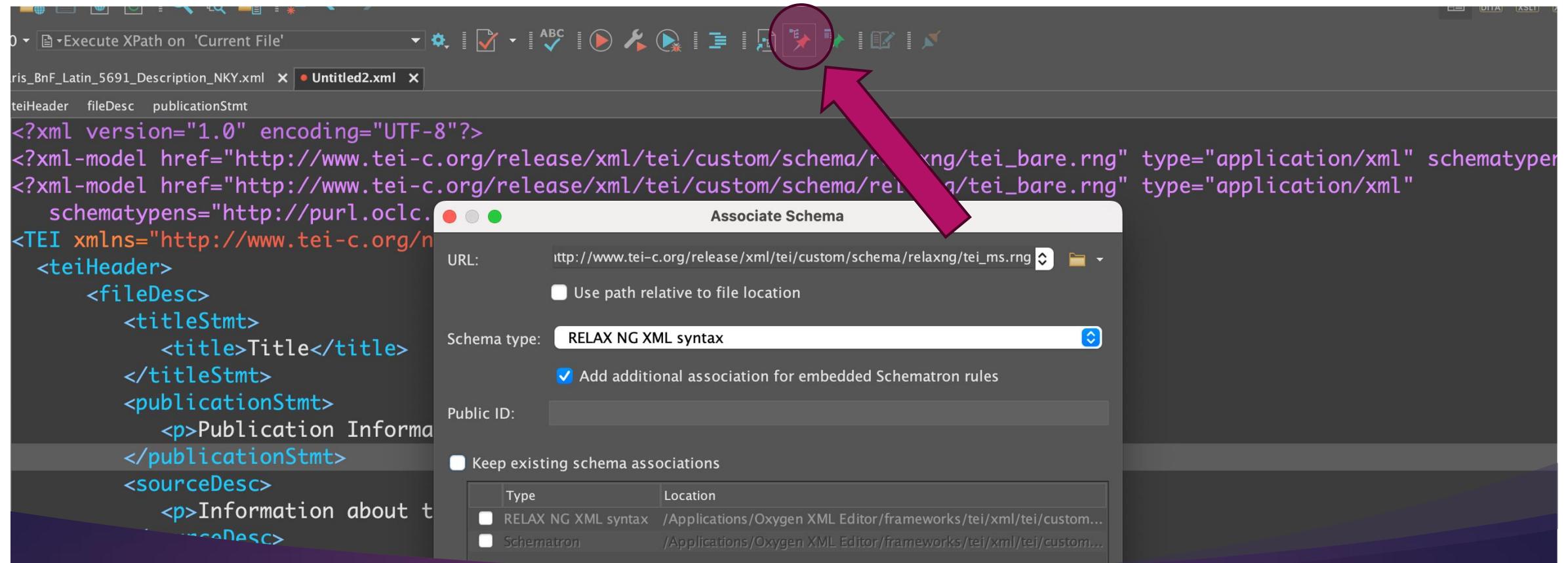
- Ex1_Paris_BnF_Latin_5691_Description_NKY.xml* ✘
- Untitled4.xml ✘
- Untitled3.xml

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_ms.rng" type="application/xml" schematypens='
3 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_ms.rng" type="application/xml"
4   schematypens="http://purl.oclc.org/dsdl/schematron"?>
5 <TEI xmlns="http://www.tei-c.org/ns/1.0">
6   <teiHeader xml:lang="en">
7     <fileDesc>
8       <titleStmt> [2 lines]
9       <publicationStmt> [15 lines]
10      <sourceDesc>
11        <msDesc xml:id="Ms-Paris-BNF-Lat-5691">
12          <msIdentifier> [5 lines]
13          <msContents> [124 lines]
14          <physDesc> [54 lines]
15          <history>
16            <origin>
17              <origDate notBefore="1390" notAfter="1400">The end of the fourteenth century.</origDate>
18              <origPlace cert="high"><name type="place">Pavia</name>, Italy.</origPlace>
19            </origin>
20            <provenance> There is an ownership mark on fol. <locus frag="42v" to="42v">42v</locus> that reads <q>Iste liber est de Angelo de Sanguineis de Roma</q>, indicating that the manuscript was owned by
21              <name type="owner">Angelo de Sanguineis de Roma</name>.</provenance>
22              <provenance>France, Paris, Bibliothèque nationale de France, Latin 8025 was also
23                ...</provenance>
```

The structure of
msDesc is well-defined
in **tei_ms** schema,
which will prevent you
from making **many**
encoding mistakes, but
NOT ALL of them

Text **Grid** **Author**

Results



Schema Association

TEI schemas

<https://www.tei-c.org/guidelines/customization/>

Customizations provided by the TEI Consortium

Customization	Description	Formats
Lite	TEI Lite, the most widely used TEI customization; includes basic elements for simple documents	ODD , DTD , RNG , XSD , HTML , PDF
TEI Tite	A constrained customization designed for use by keyboarding vendors	ODD , DTD , RNG , XSD , HTML , PDF
TEI simplePrint	An entry-level customization, focused primarily on the needs of those encoding Western European early modern printed material	ODD , DTD , RNG , XSD , HTML , PDF
jTEI Article	A highly-constrained customization intended for creating journal articles, in particular for submission to the Journal of the Text Encoding Initiative	ODD
Bare	TEI Absolutely Bare, a very barebones schema with the absolute minimum of elements	ODD , DTD , RNG , XSD
All	TEI with all modules included	ODD , DTD , RNG , XSD
Corpus	TEI for Linguistic Corpora, includes the modules for encoding linguistic corpora, ODD , DTD , RNG , XSD	
MS	TEI for Manuscript Description, includes the elements for describing manuscripts and complex physical aspects of documents	ODD , DTD , RNG , XSD
Drama	TEI with Drama, includes the TEI drama module	ODD , DTD , RNG , XSD
Speech	TEI for Speech Representation, includes the TEI module for spoken language	ODD , DTD , RNG , XSD

The following customizations use features which are not available in the DTD and XSD schema formats.

Customization	Description	Formats
Odds	TEI for authoring ODD, includes the TEI module for creating ODD files and customizations	ODD , RNG

Exercise 2

Files in Week4/Ex2

Working with the same files you used last week (Week4/Ex2)

- ▶ Change the schema association to tei_bare and check whether the file is validating
- ▶ Change the schema association to tei_all and check whether the file is validating
- ▶ Encode the following physical features of this manuscript (included under the following text in bold 'Historische Sammelhandschrift'):
 - ▶ Material of the manuscript
 - ▶ Number of leaves, including fly-leaves
 - ▶ Place of origin and Date of origin
- ▶ Use at least **three different attributes** on at least **two different elements**
- ▶ You can use the description of the Paris manuscript (Exercise 1) for inspiration.

Transcribing Primary Sources

Text (Basics Recap - week 3)

```
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
```

```
  <teiHeader>
```

```
    metadata
```

```
  </teiHeader>
```

```
  <text>
```

```
    textual content
```

```
  </text>
```

```
</TEI>
```

<text>

<front> </front>

<body> </body>

<back> </back>

</text>

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.

body (text body) contains the whole body of a single unitary text, excluding any front or back matter.

back (back matter) contains any appendixes, etc. following the main part of a text.

Div and its attributes

<div> (text division) contains a subdivision of the front, body, or back of a text.

@type characterizes the element in some sense, using any convenient classification scheme or typology.

@subtype provides a sub-categorization of the element, if needed.

@n gives a number (or other label) for an element, which is not necessarily unique within the document.

Div may contain other elements

<**p**> (paragraph) marks paragraphs in prose..

<**lg**> (line group) contains one or more verse lines functioning as a formal unit, e.g. a stanza, refrain, verse paragraph, etc.

<**l**> (verse line) contains a single, possibly incomplete, line of verse.

@met (metrical structure, conventional) contains a user-specified encoding for the conventional metrical structure of the element.

@rhyme (rhyme scheme) specifies the rhyme scheme applicable to a group of verse lines.

Div may contain milestones

<gb> (gathering beginning) marks the beginning of a new gathering or quire in a transcribed codex.

<pb> (page beginning) marks the beginning of a new page in a paginated document.

<lb> (line beginning) marks the beginning of a new (typographic) line in some edition or version of a text.

<cb> (column beginning) marks the beginning of a new column of a text on a multi-column page.

Additions, Deletions, and Omissions

<**add**> (addition) contains letters, words, or phrases inserted in the source text by an author, scribe, or a previous annotator or corrector.

<**del**> (deletion) contains a letter, word, or passage deleted, marked as deleted, or otherwise indicated as superfluous or spurious in the copy text by an author, scribe, or a previous annotator or corrector.

<**gap**> indicates a point where material has been omitted in a transcription because the material is illegible, invisible, or inaudible.

<**unclear**> contains a word, phrase, or passage which cannot be transcribed with certainty because it is illegible or inaudible in the source.

<**supplied**> signifies text supplied by the transcriber or editor for any reason; for example because the original cannot be read due to physical damage, or because of an obvious omission by the author or scribe.

Apparent Errors

<**sic**> (Latin for thus or so) contains text reproduced although apparently incorrect or inaccurate.

<**corr**> (correction) contains the correct form of a passage apparently erroneous in the copy text.

Regularization and Normalization

<**reg**> (regularization) contains a reading which has been regularized or normalized in some sense.

<**orig**> (original form) contains a reading which is marked as following the original, rather than being normalized or corrected.

Abbreviations

<abbr> (abbreviation) contains an abbreviation of any sort.

<expan> (expansion) contains the expansion of an abbreviation.

<am> (abbreviation marker) contains a sequence of letters or signs present in an abbreviation which are omitted or replaced in the expanded form of the abbreviation.

<ex> (editorial expansion) contains a sequence of letters added by an editor or transcriber when expanding an abbreviation.

Abbreviations and Expansions

- ▶ How are you doing Dr. Jones?
- ▶ <p> How are you doing <abbr>Dr.</abbr> Jones? </p>
- ▶ <p> How are you doing <abbr>Dr<am>. </am></abbr> Jones? </p>
- ▶ <p> How are you doing <expan>Doctor</expan> Jones? </p>
- ▶ <p> How are you doing <expan>D<ex>octo</ex>r</expan> Jones? </p>



Dr. *abbreviation*



- 1 (in writing) Doctor
 - *Dr. (Jane) Walker*

Source:

https://www.oxfordlearnersdictionaries.com/definition/american_english/

<choice>

- ▶ <p>

How are you doing

<choice>

<abbr>Dr<am>.</am></abbr>

<expan>D<ex>octo</ex>r</expan>

</choice>

Jones?

</p>

<choice>

<**sic**> (Latin for thus or so) contains text reproduced although apparently incorrect or inaccurate.

<**corr**> (correction) contains the correct form of a passage apparently erroneous in the copy text.

<**reg**> (regularization) contains a reading which has been regularized or normalized in some sense.

<**orig**> (original form) contains a reading which is marked as following the original, rather than being normalized or corrected.

<**unclear**> contains a word, phrase, or passage which cannot be transcribed with certainty because it is illegible or inaudible in the source.

<**supplied**> signifies text supplied by the transcriber or editor for any reason; for example because the original cannot be read due to physical damage, or because of an obvious omission by the author or scribe.

Transcription: Use of attributes

@reason gives the reason for omission or indicates why the material is hard to transcribe. Value examples: cancelled, illegible, faded, etc.

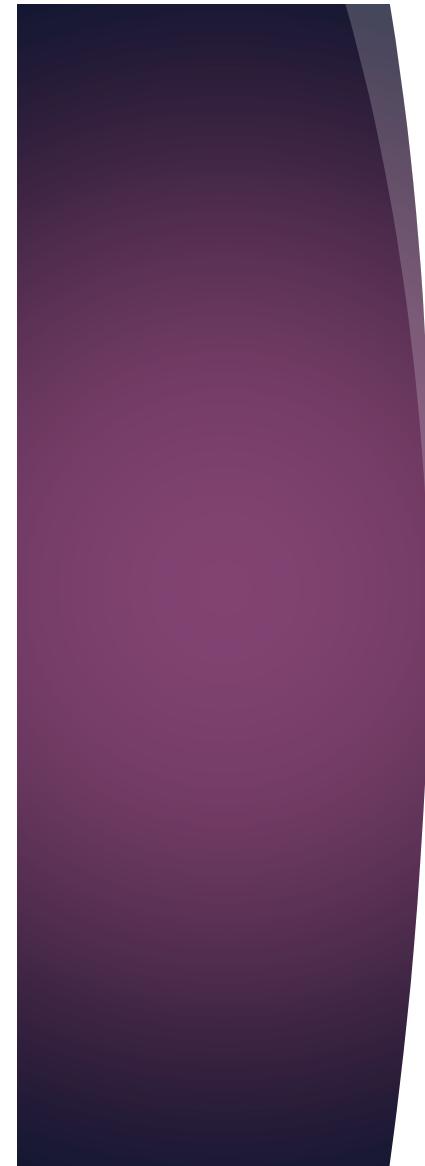
Can be used as part of **<gap>** **<unclear>** **<supplied>**

Texts in Verse: Basic Structure

```
<text>
  <body>
    <div>
      <p>
        <lg>
          <l> </l>
          <l> </l>
          <l> </l>
          <l> </l>
        </lg>
      </p>
    </div>
  </body>
</text>
```

Texts in Prose: Basic Structure

```
<text>
  <body>
    <div>
      <p>
        <choice>
          <am></am><ex></ex>
        </choice>
      </p>
    </div>
  </body>
</text>
```





The Unicode Consortium is the standards body for the internationalization of software and services. Deployed on more than 20 billion devices around the world, Unicode also provides the solution for internationalization and the architecture to support localization.

About Unicode

Technical Quick Start Guide

Support Unicode +

Adopt a Character +

Membership +

News and Events +

Emoji +

Newsletter Signup ↗

Search ...

About the Unicode Consortium

The Unicode Consortium is the standards body for the internationalization of software and services. Deployed on more than 20 billion devices around the world, Unicode also provides the solution for internationalization and the architecture to support localization.

Quick Facts

- Founded in 1988, incorporated in 1991
- Public benefit, 501(c)3 non-profit organization
- Open source standards, data, and software development
- Orchestrates the contributions of 100s of professionals, expert volunteers, and language experts
- 30+ [organizational members](#) across corporate, academic, and governmental institutions
- Funded by [membership dues](#) and [donations](#)

Operating Values

- Local solutions require global collaboration
- Localization respects and empowers users
- Interoperability across platforms serves you – and the greater good
- Transparency and open source ensure: Reliability – Security – Stability

How Did Unicode Get its Name?

The Unicode Consortium started out as the standards body for character encoding and derives its name from three main goals:

- *Universal (addressing the needs of world languages)*

Source: <https://home.unicode.org>



MUFI

Medieval Unicode Font Initiative

Home

Browse by character

Browse by code chart

Browse by range

Browse by updates

Full code chart

MUFI: The Medieval Unicode Font Initiative

Archived old MUFI site

Search the character database using the icon at the top right, or browse using the links on the left (or icon at the top left on smaller devices).

Download a snapshot of MUFI's data as: [csv](#) or [json](#) (CC BY-SA 4.0).

Disclaimer

This site is managed by scholars in Medieval studies with the aim of establishing a consensus on the use of Unicode among medievalists. It is not affiliated with or endorsed by Unicode.

Board 2016-

Tarrin Wills, Copenhagen (Chair)
Alex Speed Kjeldsen, Copenhagen (Deputy chair)
Odd Einar Haugen, Bergen
Beeke Stegmann, Copenhagen

Board 2001-2015

Odd Einar Haugen, Bergen

subsections

1. Background
2. MUFI character recommendation
3. MUFI fonts
4. Medievalist proposals to Unicode
5. LINCUA
6. Board 2001–2015
7. Other Fonts & Projects
8. Links
9. Notes

General Entities

- ▶ <!ENTITY Name "Value">
- ▶ General Entities for Special Characters
- ▶ <!DOCTYPE TEI
[
 <!ENTITY slong "ſ">
 <!ENTITY inodot "ı">
]>
- ▶ In transcription: &slong;ıc



Parameter Entities

- ▶ <!ENTITY % Name 'Value'>
- ▶ Parameter Entities for an Externally Stored List of Entities
- ▶ <!DOCTYPE TEI
 - [
 - <!ENTITY % menota_entities SYSTEM 'menota_entities.txt'>
 - %menota_entities;**
 -]>

Exercise 2: Transcription

- ▶ Associate `menota_entities.txt` with your starter XML file as a parameter entity
- ▶ Transcribe Chapter 14 of *De excidio Troiae historia* (The History of the Destruction of Troy) by Dares of Phrygia
- ▶ Use at least two different entities for special characters (you can find common entities in **MUFI_Entities_for_Hands-On_Encoding_Texts_in_Prose.pdf**)
- ▶ Use at least the following elements: `div`, `p`, `ex`, `lb`, `pb`.

MUFI: <https://mufi.info/>

MUFI Entity	Character	Name
&bar;	—	Combining abbreviation mark bar above
&barbl;	—	Combining abbreviation mark bar below
&etfin;	ȝ	Latin abbreviation sign small final et (3-shaped mark)
&er;	ȝ	Combining abbreviation mark superscript er
&ercurl;	՚	Combining abbreviation mark zigzag above curly
&qbardes;	ȝ	Latin small letter q with stroke through descender
&rsup;	ր	Combining Latin small letter r
ı	ի	Latin small letter dotless i
&inodotsup;	՚ի	Combining Latin small letter dotless i
&rum;	՞ւ	Latin abbreviation sign small rum
&	ؑ	Ampersand
&et;	ȝ	Tironian sign et
&slong;	ſ	Latin small letter long s
&rrot;	ڻ	Latin small letter r rotunda

Homework

- ▶ Prepare a single XML file for the Leipzig manuscript
- ▶ The file should validate with msDesc schema (http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_ms.rng)
- ▶ The file should contain at least the following sections correctly encoded
 - ▶ msIdentifier & msContents (week4/ex2)
 - ▶ physDesc (week5/ex1)
 - ▶ transcriptions of chapter 14 with entities (week5/ex2)
- ▶ Send your file to Katarzyna by 23:59 on Monday (3 November)

Q Go to file

t

- ✓ Encoding_Project_Portfolio_Instru...
 - Kapitan_TEI-XML_Encoding_Proj...
 - Kapitan_TEI-XML_Encoding_Proj... (selected)
 - Kapitan_TEI-XML_Encoding_Proj...
- > Week1
- > Week2
- > Week3
- > Week4

Assignment 2 (25%)

Submission Deadline: 17 November @ 23:59 (Paris time zone)

Submission Form: Folder on GitHub consisting of:

1. **ReadMe file** (Markdown, 250–300 words) describing your project, materials, and approach.
2. **Three TEI XML files** consisting of your texts encoded and validated against the TEI_all schema.
3. **Reflection note** (250–300 words) comparing your DTD design from Assignment 1 with TEI_all: what was similar, what was different, what you could and could not encode, and what you learned from this.

Detailed Instructions

1. **Select your materials and write the ReadMe file.**

- You may continue with the same source texts you identified in Assignment 1 or choose new samples of a similar length.
- If you use new samples, include unprocessed materials as well.
- Include a short AI usage statement describing any...

- ▶ [https://github.com/KAKDH/TNAH XML2025/blob/main/Encoding Project Portfolio Instructions/Kapitan TEI-XML Encoding Project Portfolio Assignment 2.pdf](https://github.com/KAKDH/TNAH_XML2025/blob/main/Encoding%20Project%20Portfolio%20Instructions/Kapitan%20TEI-XML%20Encoding%20Project%20Portfolio%20Assignment%202.pdf)
- ▶ **Due:** 17/11/2025

Portfolio Part 2