Modelling humanities data with TEI-XML

SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Dr Katarzyna Anna Kapitan
30 October 2024

<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

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

<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>

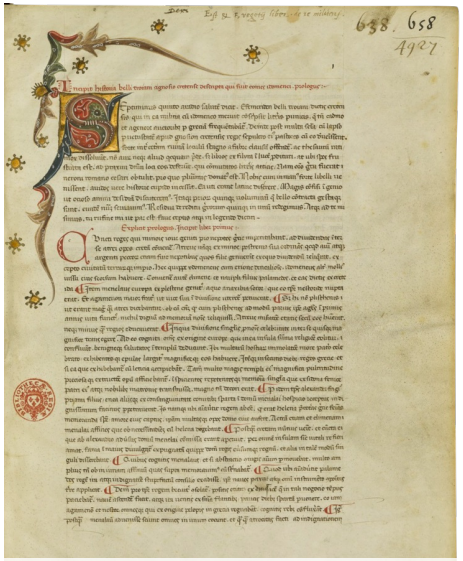
- **<objectDesc>** contains a description of the physical components making up the object which is being described.
  - **@form** a short project-specific name identifying the physical form of the carrier, for example as a codex, roll, fragment, partial leaf, cutting etc.
- **<supportDesc>** groups elements 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

- 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.

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

<bindingDesc>

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

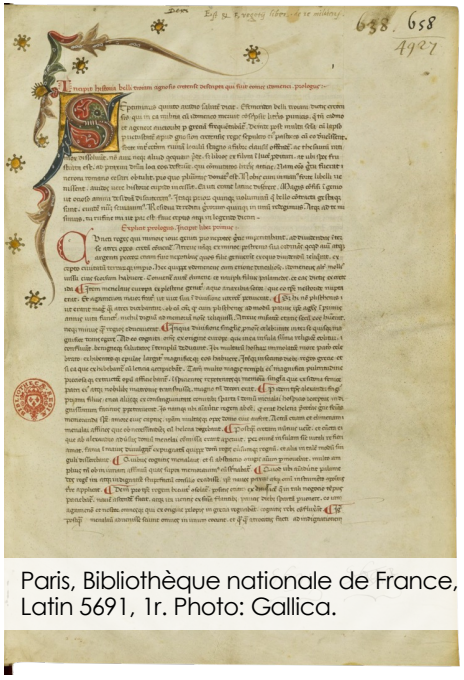
@contemporary specifies whether or not the binding is contemporary with the majority of its contents.

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

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

Exercise 1.2: Paris, BnF, Latin 5691

- Answer the following questions:
  - How many columns are there in the manuscript?
  - How many written lines are there in the manuscript?
  - How many ruled lines 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?



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.

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

<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.

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

<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.

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

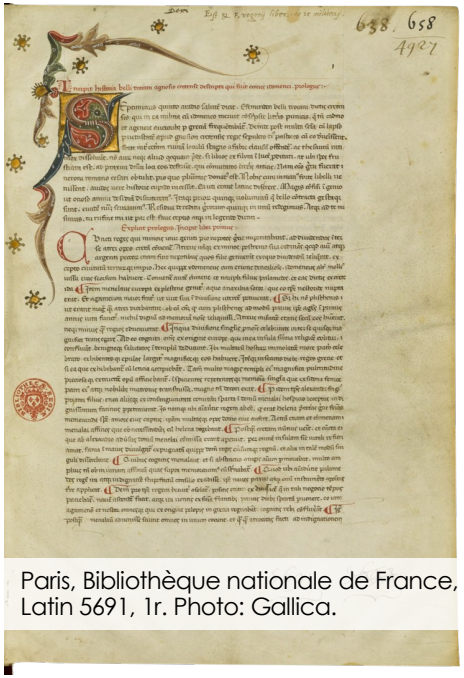
<adminInfo>

- **<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.

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

Exercise 1.3: Paris, BnF, Latin 5691

- Answer the following questions:
  - How is the binding described?
  - What is the mistake that we've made in describing the binding?
  - Who is Angelo de Sanguineis de Roma?
  - What is the mistake that we've made in describing his role?



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

Source: Kapitan & Yavuz, Manuscripts in the Digital Age: XML-Based Catalogues and Editions, ESU DH, Leipzig 2019

Kapitan, Modelling humanities data with TEI-XML

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

 (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.

X

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.

X

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.

X

Abbreviations

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

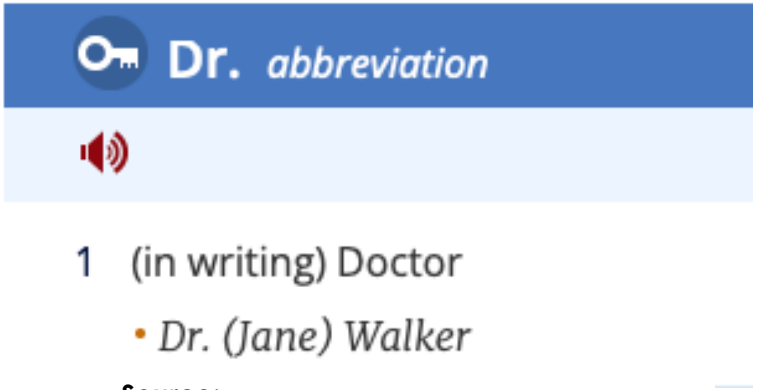
<expan> (expansion) contains the expansion of an abbreviation.

<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?



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

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

<choice>

-
  How are you doing
  <choice>

  <abbr>Dr<am>.</am></abbr>
  <expan>D<ex>octo</ex>r</expan>

  </choice>
  Jones?

<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.

X

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>

X

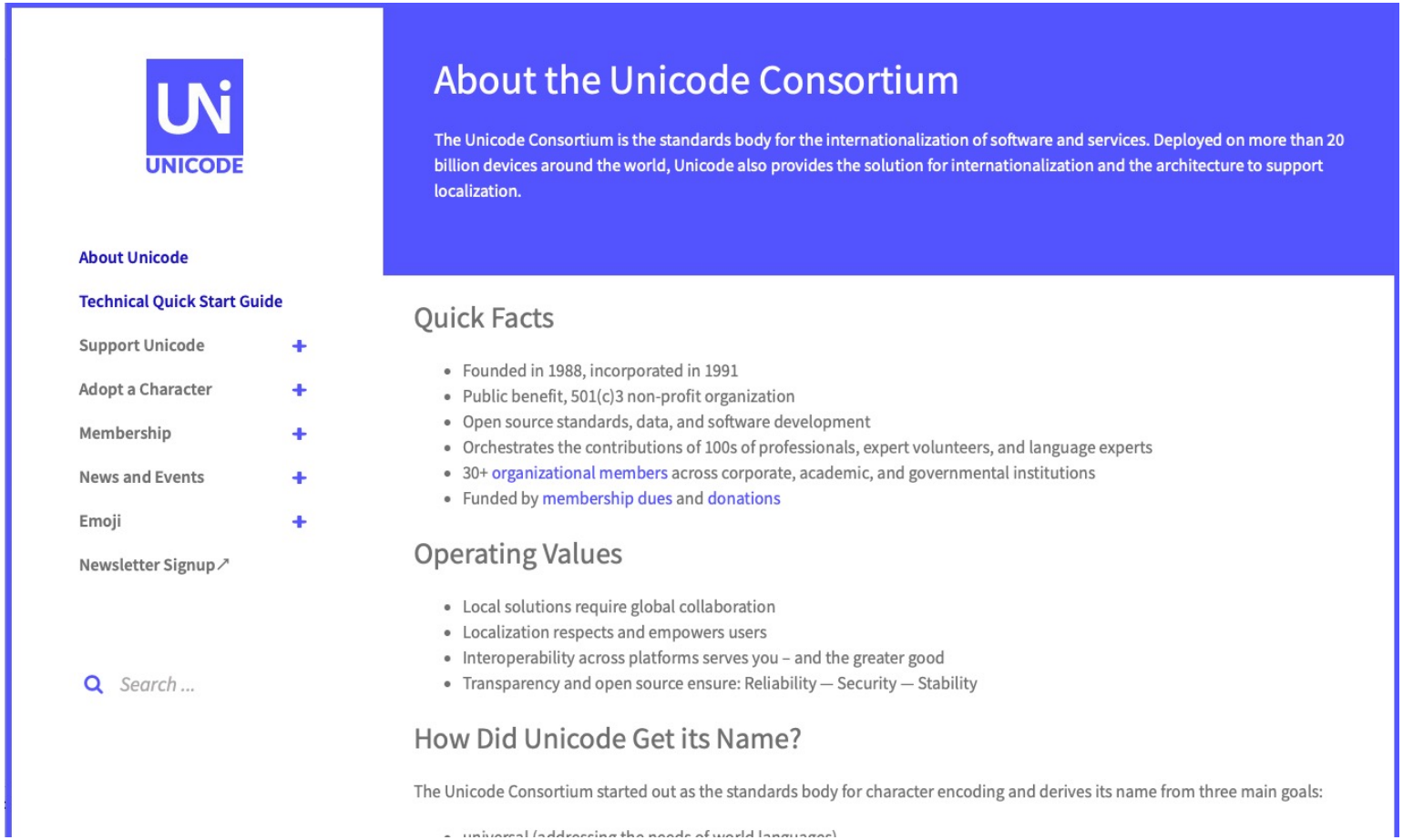
Texts in Verse: Basic Structure

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

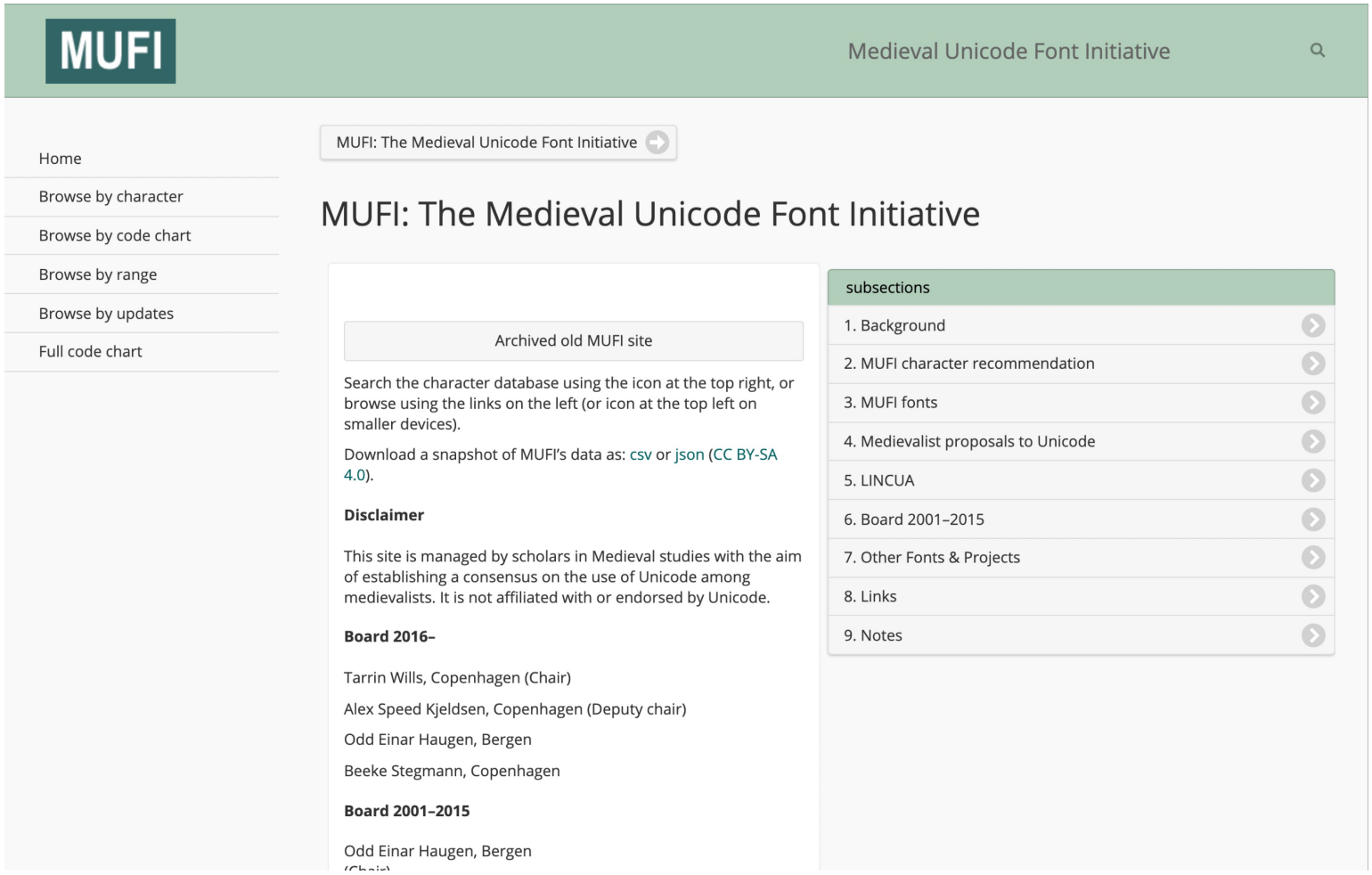
X

Texts in Prose: Basic Structure

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



Source: https://home.unicode.org



Source: https://mufi.info

Entities

- <!ENTITY Name "Value">

- <!ENTITY slong "ſ">
- <!ENTITY inodot "ı">

sic | &slong;ıc | ſic

External entity list from MENOTA:
<!ENTITY % menota_entities SYSTEM 'menota_entities.txt'> %menota_entities;

MUFI: https://mufi.info/

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

Exercise 2: Transcription

- Transcribe Chapter 14 of *De excidio Troiae historia* (The History of the Destruction of Troy) by Dares of Phrygia
- Use entities for special characters, and at least the following elements: div, p, ex, lb, pb.
- Encode chapter 14 from the following manuscripts:
  - **Group 1**: Dublin, Trinity College, IE TCD 11500
  - **Group 2:** Munich, Bayerische Staatsbibliothek, Clm 305
  - **Group 3:** St Gall, Stiftsbibliothek, Cod. Sang. 197