Modelling humanities data with TEI-XML

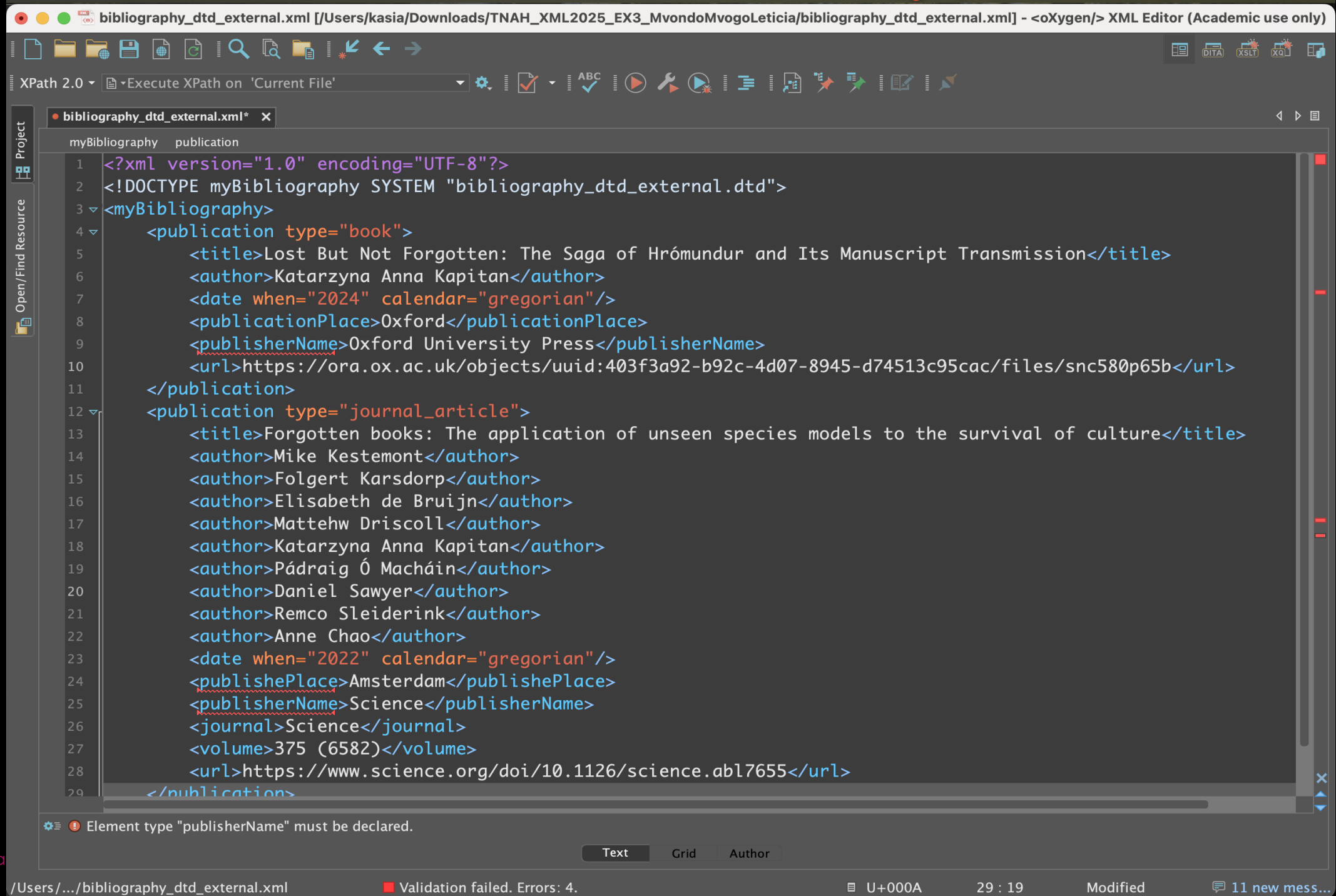
SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Dr Katarzyna Anna Kapitan
15 October 2024

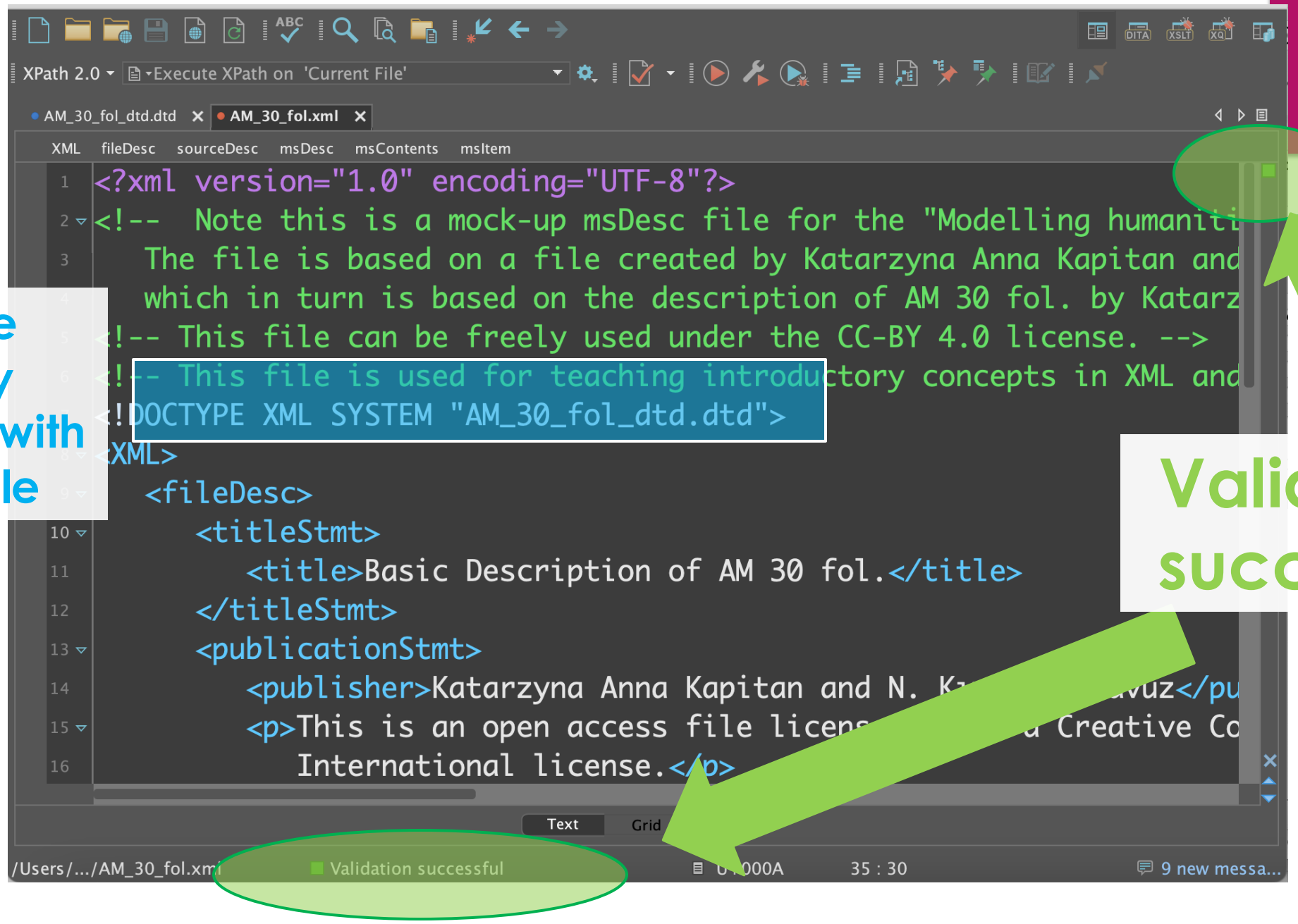
DTD Recap

- For external DTD, you need at least two files, one XML file and one DTD file.
- To associate your DTD file with your XML file you need to add the following to the beginning of the XML file:
- <!DOCTYPE *myRoot* SYSTEM "*myDTD.dtd*">

** Remember that both files must be saved in the same folder for this declaration to work.*


bibliography_dtd_external.xml [/Users/kasia/Downloads/TNAH_XML2025_EX3_MvondoMvogoLeticia/bibliography_dtd_external.xml] - <oXygen/> XML Editor (Academic use only)
XPath 2.0
Execute XPath on 'Current File'
bibliography_dtd_external.xml*
myBibliography
publication
Project
Open/Find Resource
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE myBibliography SYSTEM "bibliography_dtd_external.dtd">
<myBibliography>
<publication type="book">
<title>Lost But Not Forgotten: The Saga of Hrómundur and Its Manuscript Transmission</title>
<author>Katarzyna Anna Kapitan</author>
<date when="2024" calendar="gregorian"/>
<publicationPlace>Oxford</publicationPlace>
<publisherName>Oxford University Press</publisherName>
<url>https://ora.ox.ac.uk/objects/uuid:403f3a92-b92c-4d07-8945-d74513c95cac/files/snc580p65b</url>
</publication>
<publication type="journal_article">
<title>Forgotten books: The application of unseen species models to the survival of culture</title>
<author>Mike Kestemont</author>
<author>Folgert Karsdorp</author>
<author>Elisabeth de Bruijn</author>
<author>Mattehw Driscoll</author>
<author>Katarzyna Anna Kapitan</author>
<author>Pádraig Ó Macháin</author>
<author>Daniel Sawyer</author>
<author>Remco Sleiderink</author>
<author>Anne Chao</author>
<date when="2022" calendar="gregorian"/>
<publishePlace>Amsterdam</publishePlace>
<publisherName>Science</publisherName>
<journal>Science</journal>
<volume>375 (6582)</volume>
<url>https://www.science.org/doi/10.1126/science.abl7655</url>
Element type "publisherName" must be declared.
Text
Grid
Author
/Users/.../bibliography_dtd_external.xml
Validation failed. Errors: 4.
U+000A
29 : 19
Modified
11 new mess...

A DTD file correctly associated with the XML file



Validation successful

Find Mistakes

XML

```
<publication type="monograph_and_edition">
    <title>Forgotten books: The application of unseen species models to the survival of culture</title>
    <author>Mike Kestemont,</author>
    <author>Folgert Karsdorp,</author>
    <author>Elisabeth de Bruijn,</author>

    […]

    <date when="2022" calendar="gregorian"/>
    <url>https://www.science.org/doi/10.1126/science.abl7655</url>
    <publisherName>Science</publisherName>
    <volume>375 (6582)</volume>
  </publication>
</myBibliography>
```

Find Mistake / Missing Info

XML:

```
<publication type="journal_article">
        <title>Forgotten books: The application of unseen species models to the survival of culture</title>
        <author>Mike Kestemont</author>
        <author>Folgert Karsdorp</author>
[…]
       <date when="2022" calendar="gregorian"/>
        <journal>
            <journalTitle>Science</journalTitle>
            <volume>375 (6582)</volume>
        </journal>
        <url>https://www.science.org/doi/10.1126/science.abl7655</url>
    </publication>
```

DTD:

```
<!ELEMENT publication (title, author+, date, publisherName?, journal?, volume?, publicationPlace?, url?)>
<!ATTLIST publication type CDATA #IMPLIED>
<!ELEMENT journal (journalTitle, volume)>
<!ELEMENT journalTitle (#PCDATA)>
<!ELEMENT volume (#PCDATA)>
```

Two ways of expressing the same info: Attribute or Element?

XML:

```
<publication>
    <title>Forgotten books….</title>
    <author>Mike Kestemont</author>
    <author>Folgert Karsdorp</author>[…]
    <date when="2022" calendar="gregorian"/>
    <publisherinfo name="Science" volume="375"
number="6582"/>
</publication>
```

DTD:

```
<!ELEMENT publisherinfo EMPTY>
<!ATTLIST publisherinfo name CDATA #REQUIRED>
<!ATTLIST publisherinfo volume CDATA #IMPLIED>
<!ATTLIST publisherinfo number CDATA #IMPLIED>
```

XML:

```
<publication type="journal_article">
    <title>Forgotten books…</title>
    <author>Mike Kestemont</author>
    <author>Folgert Karsdorp</author>[…]
    <date when="2022" calendar="gregorian"/>
    <journalTitle>Science</journalTitle>
    <volume>375 (6582)</volume>
</publication>
```

DTD:

```
<!ELEMENT publication (title, author+, date,
publicationPlace?, publisherName?, journalTitle?,
volume?, url?)>
<!ATTLIST publication type (science_journal |
monograph_and_edition) #IMPLIED>
```

XML:

```
<publication type="journal_article">
      <title>Forgotten books…</title>
      <author>Mike Kestemont</author>
      <author>Folgert Karsdorp</author>
      […]
     <date when="2022" calendar="gregorian"/>
     <journalInfo>
       <journalTitle>Science</journalTitle>
       <volume>375</volume>
       <issue>6582</issue>
       <pages>1072–1074</pages>
    </journalInfo>
    <url>https://www.science.org/doi/10.1126/science.abl7655</url>
 </publication>
```

DTD:

```
<!ELEMENT publication (title, author+, date, (monographInfo | journalInfo), publicationPlace?, url?)>
<!ATTLIST publication type (monograph_and_edition | journal_article) #IMPLIED>
<!ELEMENT journalInfo (journalTitle, volume, issue?, pages?)>
<!ELEMENT journalTitle (#PCDATA)>
<!ELEMENT volume (#PCDATA)>
<!ELEMENT issue (#PCDATA)>
<!ELEMENT pages (#PCDATA)>
```

XML:

```
<publication type="journal_article">
      <title>Forgotten books…</title>
      <author>Mike Kestemont</author>
      <author>Folgert Karsdorp</author>
      […]
      <date when="2022" calendar="gregorian"/>
      <journalInfo>
        <journalTitle>Science</journalTitle>
        <volume>375</volume>
        <issue>6582</issue>
        <pages>1072–1074</pages>
      </journalInfo>
      <url>https://www.science.org/doi/10.1126/science.abl7655</url>
 </publication>
```

DTD:

```
<!ELEMENT publication (title, author+, date, (monographInfo | journalInfo), publicationPlace?, url?)>
<!ATTLIST publication type (monograph_and_edition | journal_article) #IMPLIED>
<!ELEMENT journalInfo (journalTitle, volume, issue?, pages?)>
<!ELEMENT journalTitle (#PCDATA)>
<!ELEMENT volume (#PCDATA)>
<!ELEMENT issue (#PCDATA)>
<!ELEMENT pages (#PCDATA)>
```

Text Encoding Initiative

```
<?xml version="1.0" encoding="UTF-8"?>
<workshop name="XML_workshop”>
  <instructors>
    <name>
      <firstName>Katarzyna</firstName>
      <lastName>Kapitan</lastName>
    </name>
  </instructors>
  <participants>
    <name>
      <firstName>John</firstName>
      <lastName>Doe</lastName>
    </name>
    <name>
      <firstName>Anna</firstName>
      <lastName>Smith</lastName>
    </name>
    […]
  </participants>
</workshop>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<class title="XML_class”>
  <teachers>
    <person>
      <firstName>Katarzyna</firstName>
      <lastName>Kapitan</lastName>
    </person>
  </teachers>
  <students>
    <person >
      <firstName>John</firstName>
      <lastName>Doe</lastName>
    </person>
    <person>
      <firstName>Anna</firstName>
      <lastName>Smith</lastName>
    </person>
      […]
  </students>
</class>
```

Workshop/Class/Course
Instructors/Teachers: KAK
Participants/Students: JD, AS, […]

Need for a lingua franca: TEI

Concepts

XML

Syntax

TEI

Language:
vocabulary and grammar


```
<element>
    <element attribute="value">
        content
    </element>
</element>
```

```
<p>

<note type="foot">

<head>
```

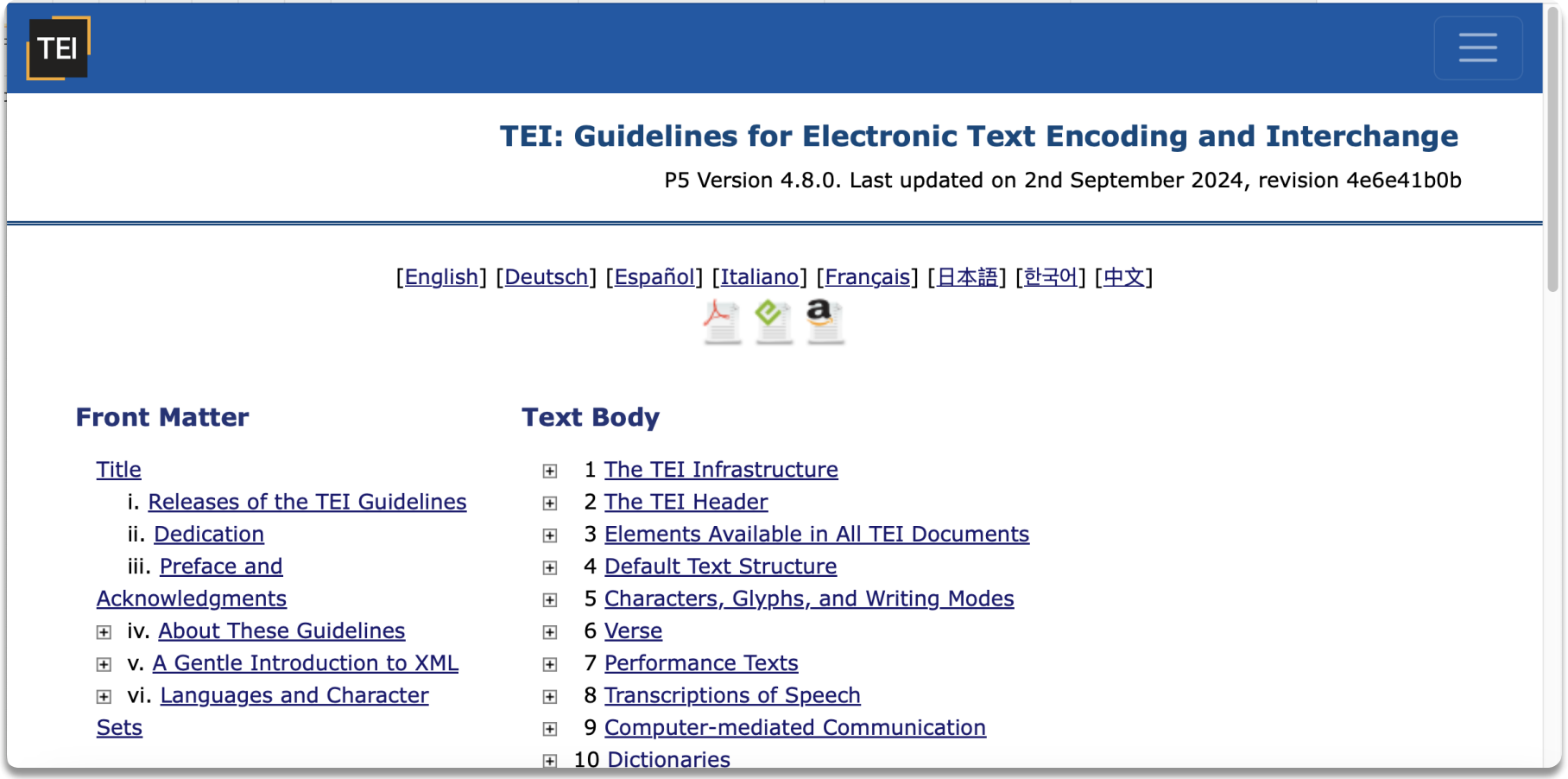
The TEI Consortium

- The TEI is an **international** and **interdisciplinary** standards project
- Established in **1987** to develop, maintain and promote hardware- and software-independent methods for encoding humanities data in electronic form.
- Since **1994**, the TEI Consortium has been issuing Guidelines which specify encoding methods for machine-readable texts.

The TEI Consortium

- Community (Membership)
- Training & Outreach
- Annual Conference
- Journal
- TEI Guidelines
- TEI Infrastructure

TEI: Guidelines for Electronic Text Encoding and Interchange



TEI Infrastructure

- Templates: https://github.com/TEIC/TEI/blob/dev/P5/Exemplars/
- Stylesheets: https://github.com/TEIC/Stylesheets
- Tools:
  - TEI Garage (convertion tool): https://teigarage.tei-c.org
  - Roma (customisation tool) (https://roma.tei-c.org)

Why to use TEI?

- International standard
- Adopted by many institutions in Europe and the USA
- Interoperable
- Machine readable
- Easy to transport/transform/reuse
- Easy to analyse

Modelling humanities data with the TEI Guidelines

The TEI Guidelines allow us to model our data so that they are:

- Sustainable
- Sharable
- Analytically rich

```xml
<front>
<!--<titlePage>書名頁</titlePag
 <divGen type="toc"/>
 <div>
  <head>序</head>
  <p> ... </p>
 </div>
</front>
```

```xml
<front>
 <divGen type="toc"/>
 <div>
  <head>Préface</head>
  <p> ... </p>
 </div>
</front>
```

```xml
<front>
<!--<titlePage>...</titlePage>-->
 <divGen type="toc"/>
 <div>
  <head>Preface</head>
  <p> ... </p>
 </div>
</front>
```

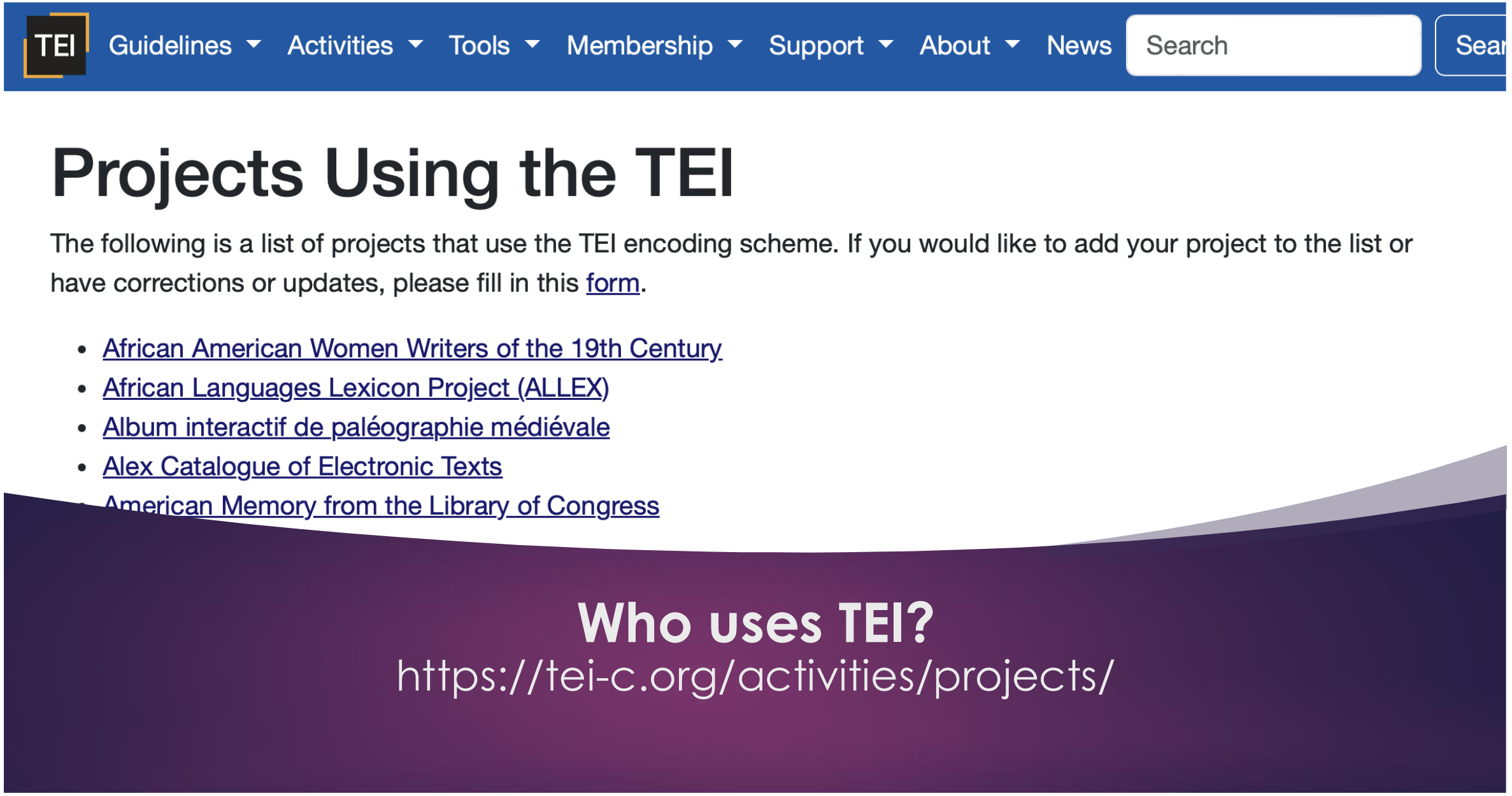
Source: https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-div.html

```xml
<ab>
 <lb n="1"/>  بسم الله الـ<supplied reason="lost">ـرّ</supplied>حمن الرّحيم
 <lb n="2"/>  هذ<supplied reason="lost">ا</supplied> ـم<supplied reason="lost">ـا</supplied>
 <lb n="3"/>  <supplied reason="lost">ـع</supplied><unclear>_</unclear><supplied reason="los
 <lb n="4"/>  فعلي أبو داره يدفع إلي مقاره سلمون وأتناس
 <lb n="5"/>  فإن لم يجدهم فعلي أبو داره ما كان عليهم من شي
 <lb n="6"/>  شهد على ذلك أيّوب مولى أيّان بن عاصم
 <lb n="7"/>  وعطا بن توّابة اللخمي وسعيد مولي أبي إبراهيم
 <lb n="8"/>  وإبراهيم بن سالم البختي وسالم بن إسحق البختي وعبيد بن <unclear>مجناد</unclear>
</ab>
```

Source: https://arabic-tei-workshop.github.io/index.html

What to use TEI for?

- Manuscript descriptions
- Transcriptions of texts
- Editions of texts (documentary, genetic, critical)
- Annotated corpora (linguistic, spatial, prosopography)
- Dictionaries
- etc.

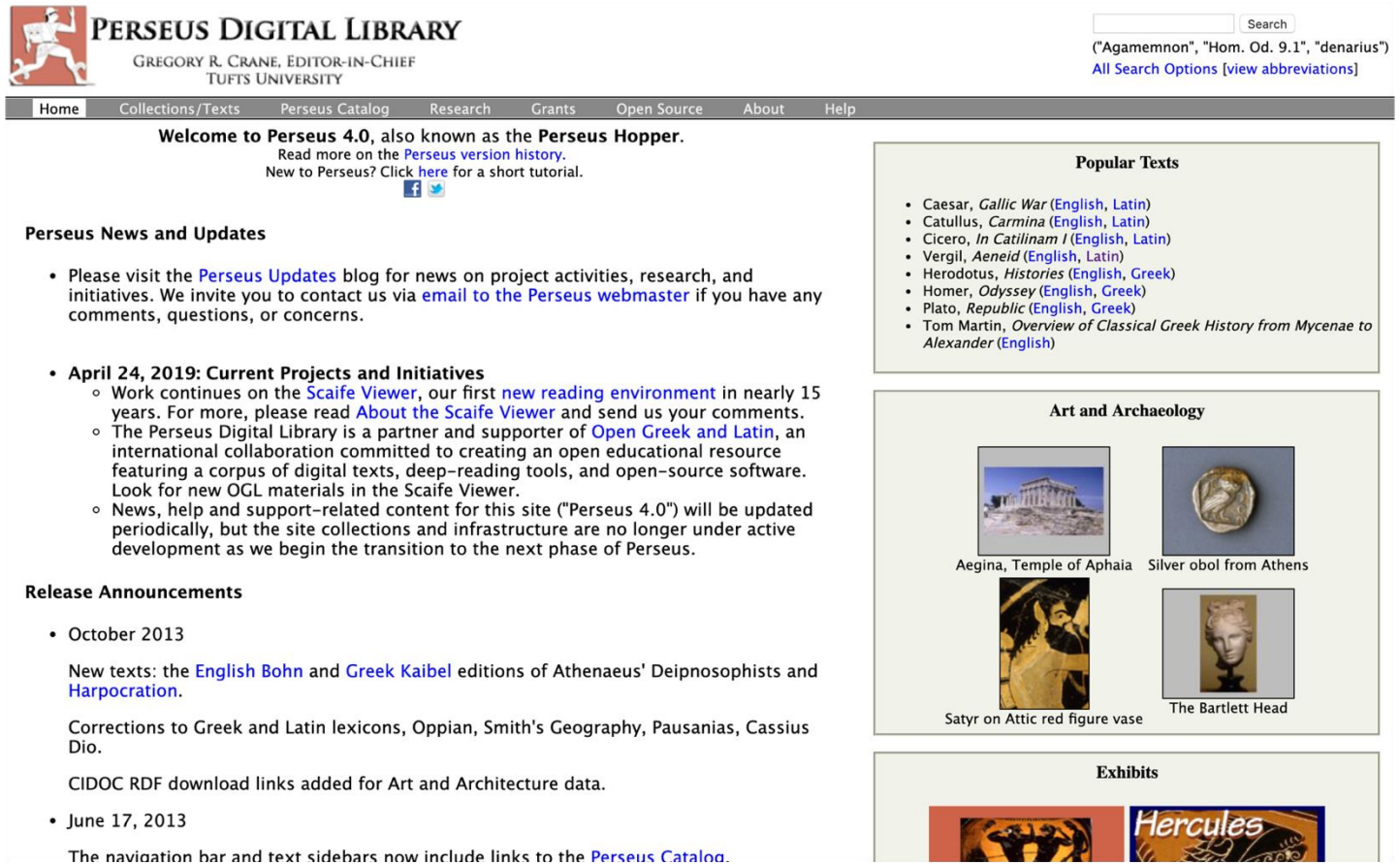


Who uses TEI?

https://tei-c.org/activities/projects/

TEI for Text Editing

Perseus Digital Library



http://www.perseus.tufts.edu

Medieval Nordic Text Archive

Medieval Nordic Text Archive
www.menota.org

Arkiv for nordiske middelaldertekster
Arkiv för nordiska medeltidstexter
Safn norrænna miðaldatexta

Norsk

Background

Menota is a network of leading Nordic archives, libraries and research departments working with medieval texts and manuscript facsimiles. The aim of Menota is to preserve and publish medieval texts in digital form and to adapt and develop encoding standards necessary for this work. The archive will contain texts in the Nordic languages as well as in Latin. There are now 15 members of Menota (see bottom of the page) and new members are welcome to join the network. Since its founding in 2001, the archive has been led by Odd Einar Haugen at the University of Bergen

News (in Norwegian)

- Konungs skuggsjá ferdig annotert (19.09.2018).
- Menota på YouTube (16.08.2018).
- Egils saga Skalla-Grímssonar i Wolfenbüttelbók i arkivet (22.06.2018).
- Kalkars ordbok er på nettet med fulle søkemuligheter (31.03.2017).
- Fjorten nye tekster i arkivet (27.07.2016).
- Avtale mellom Menota og Universitetsbiblioteket i Bergen (02.06.2016).

Text archive

Menota can now offer over 40 Medieval Nordic texts (around 1.7 million words), several of which are fully lemmatised. The majority of the texts are Old Icelandic or Old Norwegian, but there are also some Old Swedish texts. Old Danish texts as well as Latin ones (from the Nordic countries) are most welcome. A catalogue with full search facilities was opened on 29 June 2007.

The texts have been encoded on one or more levels. The most widely used level is the diplomatic level (as seen in many Arnamagnæan editions), while some texts have also been encoded on a facsimile level (i.e. in a very close transcription) and some on a normalised level (as in the Íslenzk fornrit series). These levels are specified under Facs, Dipl and Norm in the catalogue. By clicking the file name in the second column of the catalogue, you will be able to read the text on one or more levels, depending on how many levels have been encoded.

Foundation

Menota was established on 10 September 2001 at a meeting in Oslo between major Nordic institutions.

- Initial meeting, Oslo 10 September 2001

Statutes

The Menota statutes were passed at the council meeting in Reykjavík on 6 September 2003.

- The Menota statutes (Norwegian text)

Council

The council is an advisory body in which each participating institution is represented by one member. Different departments or sections at the same university are recognised as individual institutions. In addition to the permanent members, other representatives may join the meeting at the discretion of the board.

- Menota's council 2019-2021
- Menota's council 2016-2018
- Menota's council 2013-2015
- Menota's council 2010-2012
- Menota's council 2007-2009
- Menota's council 2004-2006
- Menota's council 2002-2003

Notice of meeting

- Council meeting, Bergen 26.08.2019

https://menota.org

BECKETT ARCHIVE

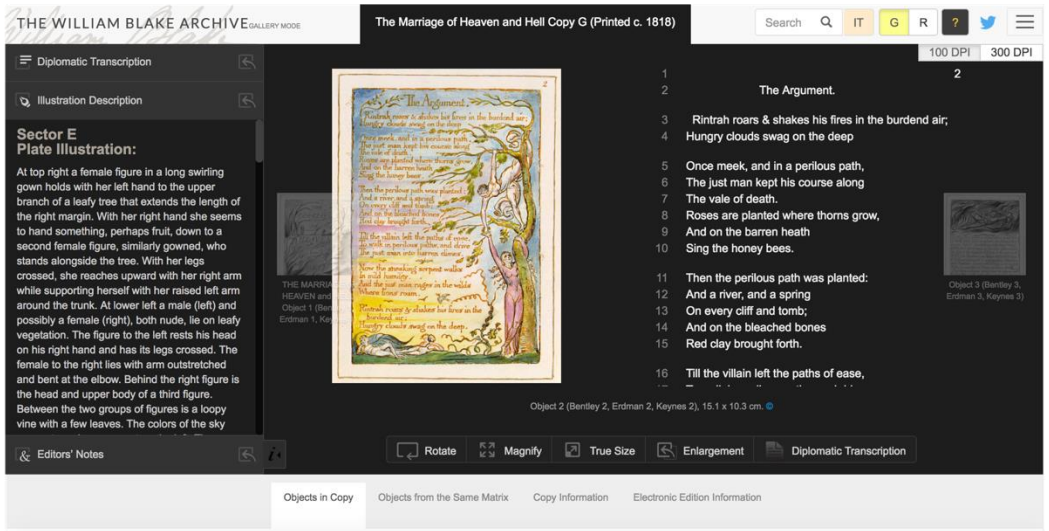
www.beckettarchive.org

And many more

- Base de Français Médiéval (https://txm-bfm.huma-num.fr)
- Le mariage sous L'Ancien Régime (https://mariage.uvic.ca)
- Newton Manuscript Project (https://www.newtonproject.ox.ac.uk/)
- etc.

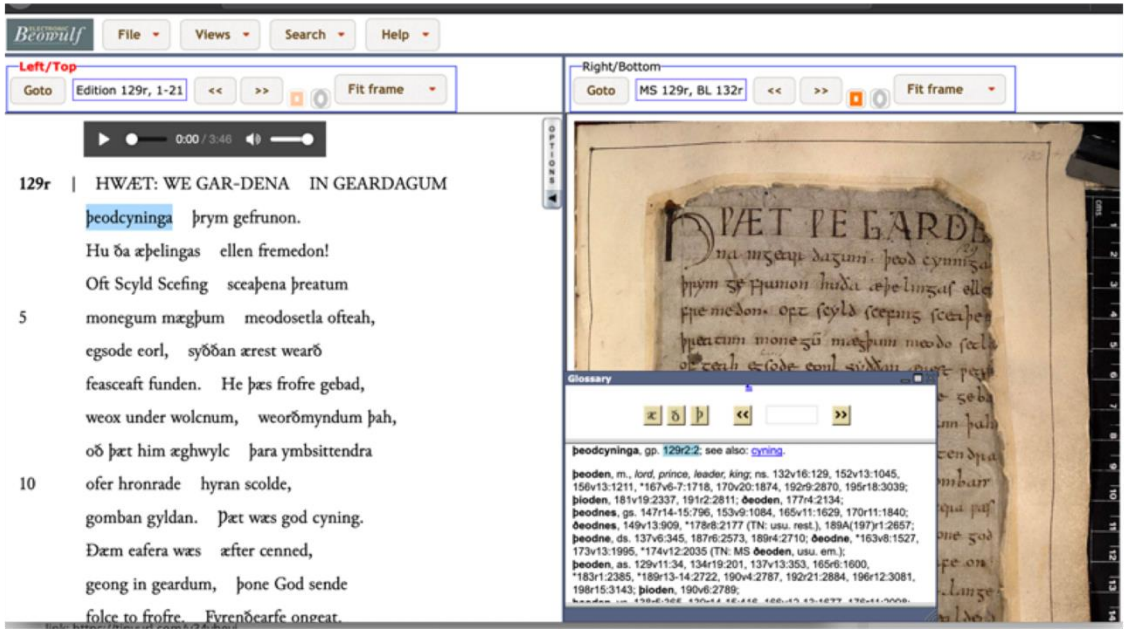
But not everyone

The William Blake Archive



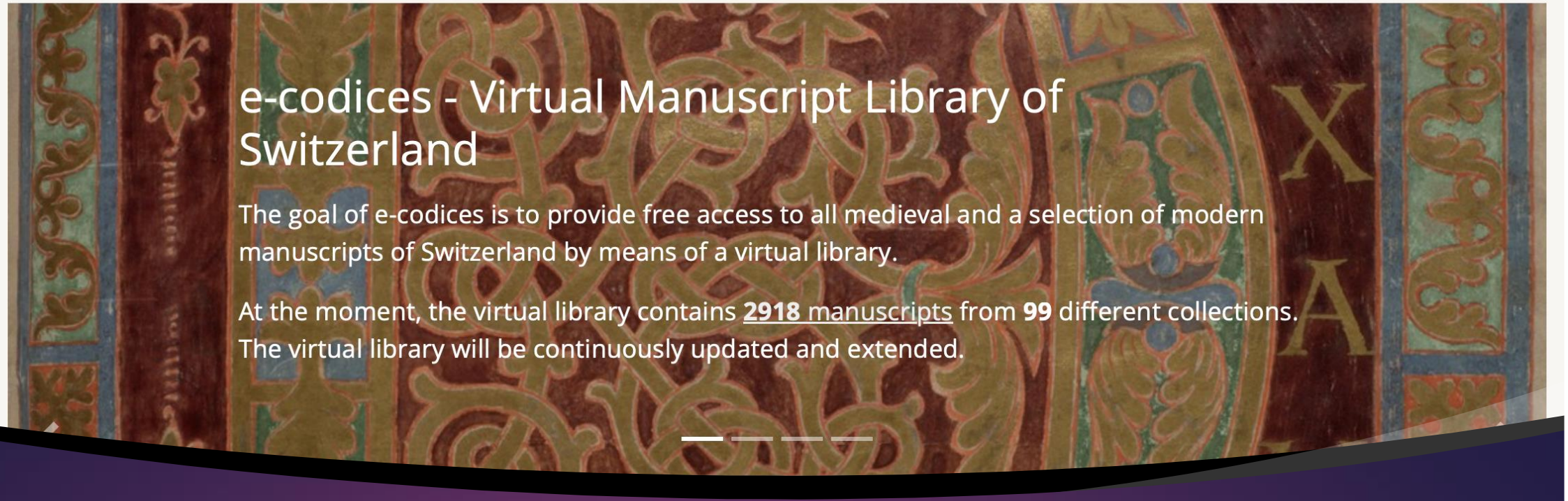
http://www.blakearchive.org

Electronic Beowulf



https://ebeowulf.uky.edu

TEI for Manuscript Cataloguing

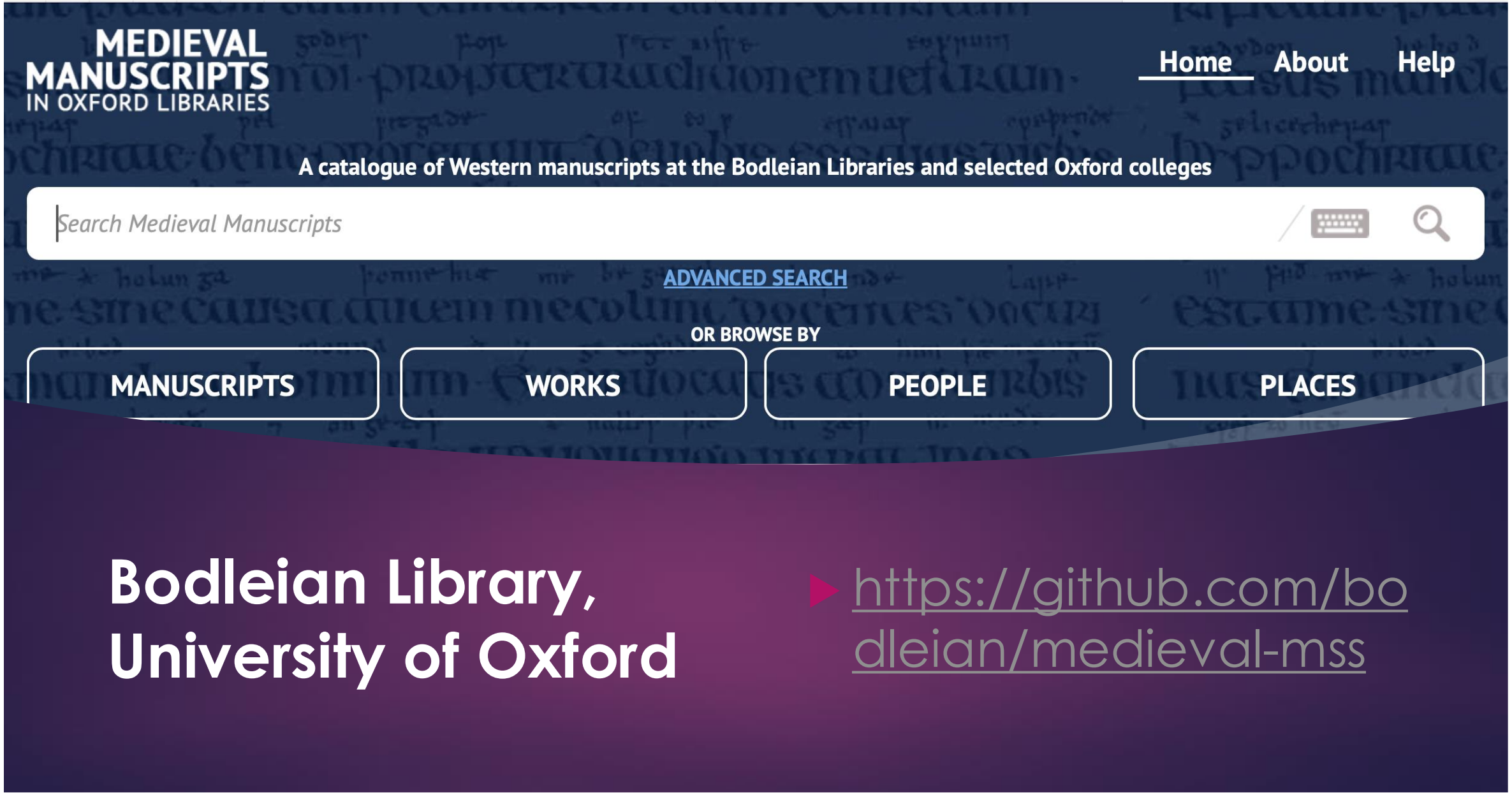


The Virtual Library of Switzerland: e-codices

- https://www.e-codices.ch/en/about/metadata

University of Pennsylvania Libraries: OPenn

HTTPS://GITHUB.COM/SIMS-MSS/OPENN-XML



Bodleian Library, University of Oxford

- https://github.com/bodleian/medieval-mss

And many more

- FIHRIST, a union Catalogue of Manuscripts from the Islamicate World (https://www.fihrist.org.uk)
- Handrit, a union Catalogue of Icelandic manuscripts (https://handrit.is)
- Cambridge Digital Library (https://cudl.lib.cam.ac.uk)
- etc.

But not everyone

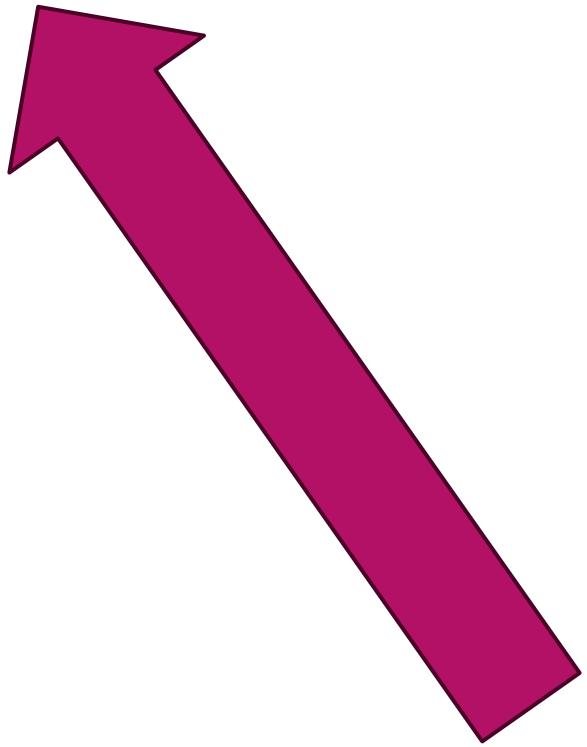
- Many archives and libraries use **MARC (machine-readable cataloguing) standard**, maintained by the Library of Congress to create machine-readable description of items catalogued by libraries.
  - See: https://www.loc.gov/marc/bibliographic/
- Many archives and libraries (notably Calames and BnF) use **Encoded Archival Description (EAD)**, an XML standard for encoding archival finding aids, maintained by the Society of American Archivists.
  - EAD Documentation: https://github.com/SAA-SDT/EAD3

TEI Structure

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

    <teiHeader>
        metadata
    </teiHeader>
    <text>
        textual content
    </text>

</TEI>
```



The root element of a TEI file is TEI (not XML) & it contains an attribute which specifies the TEI namespace

teiHeader

- **fileDesc** (file description) contains a full bibliographic description of an electronic file
- **encodingDesc** (encoding description) documents the relationship between an electronic text and the source or sources from which it was derived.
- **profileDesc** (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.
- **revisionDesc** (revision description) summarizes the revision history for a file.

fileDesc

```
<fileDesc>
  <titleStmt>
    <title> [...] </title>
  </titleStmt>
</fileDesc>
```

- **titleStmt** (title statement) groups information about the title of a work and those responsible for its content

fileDesc

```
<fileDesc>

  <publicationStmt>
    <p> [...] </p>
  </publicationStmt>

</fileDesc>
```

- **publicationStmt** (publication statement) groups information concerning the publication or distribution of an electronic or other text.

fileDesc

```
<fileDesc>
  <sourceDesc>
    <p> [...] </p>
  </sourceDesc>
</fileDesc>
```

- **sourceDesc**(source description) describes the source(s) from which an electronic text was derived or generated (a bibliographic description in the case of a digitized text, or a phrase ‘born digital’).

Example: teiHeader

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Thomas Paine: Common sense, a
       machine-readable transcript</title>
   <respStmt>
    <resp>compiled by</resp>
    <name>Jon K Adams</name>
   </respStmt>
  </titleStmt>
```

Source: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD7

Example: teiHeader

```
<publicationStmt>
 <distributor>Oxford Text Archive.</distributor>
 <address>
  <addrLine>Oxford University Computing Services,</addrLine>
  <addrLine>13 Banbury Road,</addrLine>
  <addrLine>Oxford OX2 6RB,</addrLine>
  <addrLine>UK</addrLine>
 </address>
</publicationStmt>
```

Source: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD7

```
<sourceDesc>
 <biblStruct>
  <monogr>
   <editor>Foner, Philip S.</editor>
   <title>The collected writings of Thomas Paine</title>
   <imprint>
    <pubPlace>New York</pubPlace>
    <publisher>Citadel Press</publisher>
    <date>1945</date>
   </imprint>
  </monogr>
 </biblStruct>
</sourceDesc>
```

Source: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD7

```
<encodingDesc>
 <samplingDecl>
  <p>Editorial notes in the Foner edition have not
      been reproduced. </p>
  <p>Blank lines and multiple blank spaces, including paragraph
      indents, have not been preserved. </p>
 </samplingDecl>
 <editorialDecl>
  <correction status="high"
   method="silent">
   <p>The following errors
       in the Foner edition have been corrected:
   <list>
     <item>p. 13 l. 7 cotemporaries contemporaries</item>
     <item>p. 28 l. 26 [comma] [period]</item>
     <item>p. 84 l. 4 kin kind</item>
     <item>p. 95 l. 1 stuggle struggle</item>
     <item>p. 101 l. 4 certainy certainty</item>
     <item>p. 167 l. 6 than that</item>
     <item>p. 209 l. 24 publshed published</item>
    </list>
   </p>
  </correction>
```

Source: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD7

```
<profileDesc>
 <creation>
  <date>1774</date>
 </creation>
 <langUsage>
  <language ident="en" usage="100">English.</language>
 </langUsage>
 <textClass>
  <keywords scheme="#lcsh">
   <term>Political science</term>
   <term>United States -- Politics and government -
        Revolution, 1775-1783</term>
  </keywords>
  <classCode scheme="#lc">JC 177</classCode>
 </textClass>
</profileDesc>
```

Source: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD7

```
<revisionDesc>
 <change when="1996-01-22" who="#MSM"> finished proofreading </change>
 <change when="1995-10-30" who="#LB"> finished proofreading </change>
 <change notBefore="1995-07-04" who="#RG"> finished data entry at end of term </change>
 <change notAfter="1995-01-01" who="#RG"> began data entry before New Year 1995 </change>
</revisionDesc>
```

Source: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD7

Exercise 1: Ex1_teiHeader.xml

1. Analyse **Ex1_teiHeader.xml** and find information about
   - the language and the author of the text encoded in this document
   - the language and the author (creator) of this XML document
2. Analyse **Ex1_teiHeader.xml** and find all the dates in the document
   - How are they encoded? Is it clear what the dates refer to?
     - *If yes, what makes it clear?*
     - *If not, how could the encoding of the dates be improved?*
3. Using the online version of the TEI Guidelines find out how to:
   - Encode information about about the license under which the file can be distributed
   - Add this element in the correct place in the document

Text

```
<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.

Div: Example

```
<div type="letter">
 <opener>
  <dateline>
   <placeName>Rimaone</placeName>
   <date when="2006-11-21">21 Nov 06</date>
  </dateline>
  <salute>Dear Susan,</salute>
 </opener>
 <p>Thank you very much for the assistance splitting those
   logs. I'm sorry about the misunderstanding as to the size of
   the task. I really was not asking for help, only to borrow the
   axe. Hope you had fun in any case.</p>
 <closer>
  <salute>Sincerely yours,</salute>
  <signed>Seymour</signed>
 </closer>
 <postscript>
  <label>P.S.</label>
  <p>The collision occured on <date when="2001-07-06">06 Jul 01</date>.</p>
 </postscript>
</div>
```

Exercise 2:
The First Folio of Shakespeare's plays

- Open Ex2_Text.txt and Ex2_Image.jpg (image & text of page 101 of the First Folio of Shakespeare's plays, Bodleian Arch. G c.7)
- In Oxygen Editor create a new TEI document using the TEI_All template.
- Encode all relevant metadata in the **teiHeader** element.
- Encode the text in the **text** element.
- **At home:**
  - Compare your file with the official encoding of the First Folio of Shakespeare's plays and Katarzyna's encoding (on GitHuB)
  - See also: https://firstfolio.bodleian.ox.ac.uk/download/xml/F-ado.xml