



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' ABC

Project myBibliography publication

```
<?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>
    </publication>

```

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 screenshot of the Oxygen XML Editor interface. The top menu bar includes icons for file operations (New, Open, Save, Find, Replace, etc.) and document types (DITA, XSLT, XQ, etc.). The toolbar below has buttons for XPath 2.0, Execute XPath on 'Current File', and various editing tools. The main window shows an XML file named 'AM_30_fol.xml' and its associated DTD file 'AM_30_fol_dtd.dtd'. The XML code includes a DOCTYPE declaration pointing to the DTD file. A large green arrow points from the text 'Validation successful' in a white box on the right towards the validation status bar at the bottom. A green oval highlights the status bar message 'Validation successful'. Another green oval highlights the status bar message '9 new mess...'. The status bar also displays the file path '/Users/.../AM_30_fol.xml', the page number '35 : 30', and a message count '9 new mess...'. The left margin of the editor has line numbers 1 through 16.

A DTD file correctly associated with the XML file

Validation successful

```
<?xml version="1.0" encoding="UTF-8"?>




<!DOCTYPE XML SYSTEM "AM_30_fol_dtd.dtd">
<XML>
<fileDesc>
    <titleStmt>
        <title>Basic Description of AM 30 fol.</title>
    </titleStmt>
    <publicationStmt>
        <publisher>Katarzyna Anna Kapitan and N. K. Kavuz</publisher>
        <p>This is an open access file licensed under a Creative Commons International license.</p>
    </publicationStmt>
</fileDesc>
```

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>
```

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

```
<?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>
```

Need for a lingua franca: TEI



XML

TEI

Concepts



Syntax

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

**Language:
vocabulary and grammar**

```
<p>
<note type="foot">
<head>
```

© 2007 Syd Bauman, Julia Flanders, and the Women Writers Project. Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

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.



Text
Encoding
Initiative

The TEI Consortium

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

TEI: Guidelines for Electronic Text Encoding and Interchange

The screenshot shows the homepage of the TEI Guidelines website. At the top left is a 'TEI' logo, and at the top right is a three-line menu icon. The main title 'TEI: Guidelines for Electronic Text Encoding and Interchange' is centered above a subtitle 'P5 Version 4.8.0. Last updated on 2nd September 2024, revision 4e6e41b0b'. Below the title are language links for English, Deutsch, Español, Italiano, Français, 日本語, 한국어, and 中文, each accompanied by a small document icon. The page is divided into two main sections: 'Front Matter' on the left and 'Text Body' on the right. The 'Front Matter' section includes links for 'Title', 'Releases of the TEI Guidelines', 'Dedication', 'Preface and Acknowledgments', 'About These Guidelines', 'A Gentle Introduction to XML', and 'Languages and Character Sets'. The 'Text Body' section lists chapters 1 through 10: 'The TEI Infrastructure', 'The TEI Header', 'Elements Available in All TEI Documents', 'Default Text Structure', 'Characters, Glyphs, and Writing Modes', 'Verse', 'Performance Texts', 'Transcriptions of Speech', 'Computer-mediated Communication', and 'Dictionaries'. Each chapter link is preceded by a plus sign.

Front Matter

- [Title](#)
 - i. [Releases of the TEI Guidelines](#)
 - ii. [Dedication](#)
 - iii. [Preface and Acknowledgments](#)
- [About These Guidelines](#)
- [A Gentle Introduction to XML](#)
- [Languages and Character Sets](#)

Text Body

- + 1 [The TEI Infrastructure](#)
- + 2 [The TEI Header](#)
- + 3 [Elements Available in All TEI Documents](#)
- + 4 [Default Text Structure](#)
- + 5 [Characters, Glyphs, and Writing Modes](#)
- + 6 [Verse](#)
- + 7 [Performance Texts](#)
- + 8 [Transcriptions of Speech](#)
- + 9 [Computer-mediated Communication](#)
- + 10 [Dictionaries](#)

TEI Infrastructure

- ▶ Templates:
<https://github.com/TEIC/TEI/blob/dev/P5/Exemplars/>
- ▶ Stylesheets:
<https://github.com/TEIC/Stylesheets>
- ▶ Tools:
 - ▶ TEI Garage (conversion 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



Text
Encoding
Initiative

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

- Sustainable
- Sharable
- Analytically rich

```

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

```

```

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

```

```

<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>

```

<ab>
<lb n="1"/> حمْن الرَّحِيم بِسْم اللَّهِ الرَّحْمَنِ الرَّحِيمِ
<lb n="2"/> هذَا مَا مُنْكَرٌ عَلَيْهِ مِنْ شَيْءٍ
<lb n="3"/> فَعَلَى أَبْوَدَارِهِ يَدْفَعُ إِلَى مَقَارِهِ سَلْمُونُ وَأَتْنَاسُ
<lb n="4"/> فَإِنْ لَمْ يَجِدْهُمْ فَعْلِي أَبْوَدَارِهِ مَا كَانَ عَلَيْهِمْ مِنْ شَيْءٍ
<lb n="5"/> شَهَدَ عَلَى ذَلِكَ أَيُّوبُ مُولَى أَيَّانَ بْنَ عَاصِمٍ
<lb n="6"/> وَعَطَا بْنَ تَوَابَةَ الْلَّخْمِيَّ وَسَعِيدَ مُولَى أَبْيَ إِبْرَاهِيمَ
<lb n="7"/> وَإِبْرَاهِيمَ بْنَ سَالِمَ الْبَخْتِيَّ وَسَالِمَ بْنَ إِسْحَاقَ الْبَخْتِيَّ وَعَبِيدَ بْنَ
<lb n="8"/> مُجَنَّادَ وَإِبْرَاهِيمَ بْنَ سَالِمَ الْبَخْتِيَّ وَسَالِمَ بْنَ إِسْحَاقَ الْبَخْتِيَّ وَعَبِيدَ بْنَ
</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.

Projects Using the TEI

The following is a list of projects that use the TEI encoding scheme. If you would like to add your project to the list or have corrections or updates, please fill in this [form](#).

- [African American Women Writers of the 19th Century](#)
- [African Languages Lexicon Project \(ALLEX\)](#)
- [Album interactif de paléographie médiévale](#)
- [Alex Catalogue of Electronic Texts](#)
- [American Memory from the Library of Congress](#)

Who uses TEI?

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

TEI for Text Editing

PERSEUS DIGITAL LIBRARY



PERSEUS DIGITAL LIBRARY

GREGORY R. CRANE, EDITOR-IN-CHIEF
TUFTS UNIVERSITY

[Home](#) [Collections/Texts](#) [Perseus Catalog](#) [Research](#) [Grants](#) [Open Source](#) [About](#) [Help](#)

Welcome to Perseus 4.0, also known as the **Perseus Hopper**.

Read more on the [Perseus version history](#).
New to Perseus? Click [here](#) for a short tutorial.



Perseus News and Updates

- Please visit the [Perseus Updates](#) blog for news on project activities, research, and initiatives. We invite you to contact us via [email to the Perseus webmaster](#) if you have any comments, questions, or concerns.
- **April 24, 2019: Current Projects and Initiatives**
 - Work continues on the [Scaife Viewer](#), our first [new reading environment](#) in nearly 15 years. For more, please read [About the Scaife Viewer](#) and send us your comments.
 - The Perseus Digital Library is a partner and supporter of [Open Greek and Latin](#), an international collaboration committed to creating an open educational resource featuring a corpus of digital texts, deep-reading tools, and open-source software. Look for new OGL materials in the Scaife Viewer.
 - News, help and support-related content for this site ("Perseus 4.0") will be updated periodically, but the site collections and infrastructure are no longer under active development as we begin the transition to the next phase of Perseus.

Release Announcements

- October 2013
New texts: the [English Bohn](#) and [Greek Kaibel](#) editions of Athenaeus' *Deipnosophists* and *Harpocration*.
- Corrections to Greek and Latin lexicons, Oppian, Smith's Geography, Pausanias, Cassius Dio.
- CIDOC RDF download links added for Art and Architecture data.
- June 17, 2013
The navigation bar and text sidebars now include links to the [Perseus Catalog](#).

("Agamemnon", "Hom. Od. 9.1", "denarius")
[All Search Options](#) [view abbreviations]

Popular Texts

- Caesar, *Gallic War* ([English](#), [Latin](#))
- Catullus, *Carmina* ([English](#), [Latin](#))
- Cicero, *In Catilinam I* ([English](#), [Latin](#))
- Vergil, *Aeneid* ([English](#), [Latin](#))
- Herodotus, *Histories* ([English](#), [Greek](#))
- Homer, *Odyssey* ([English](#), [Greek](#))
- Plato, *Republic* ([English](#), [Greek](#))
- Tom Martin, *Overview of Classical Greek History from Mycenae to Alexander* ([English](#))

Art and Archaeology



Aegina, Temple of Aphaia



Silver obol from Athens



Satyr on Attic red figure vase



The Bartlett Head

Exhibits



[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-Grimssonar 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 Íslensk 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

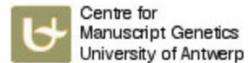
- ▶ Council meeting, Bergen 26.08.2019

<https://menota.org>

BECKETT ARCHIVE

SAMUEL BECKETT
DIGITAL MANUSCRIPT PROJECT

KRAPP'S LAST TAPE / LA DERNIÈRE BANDE

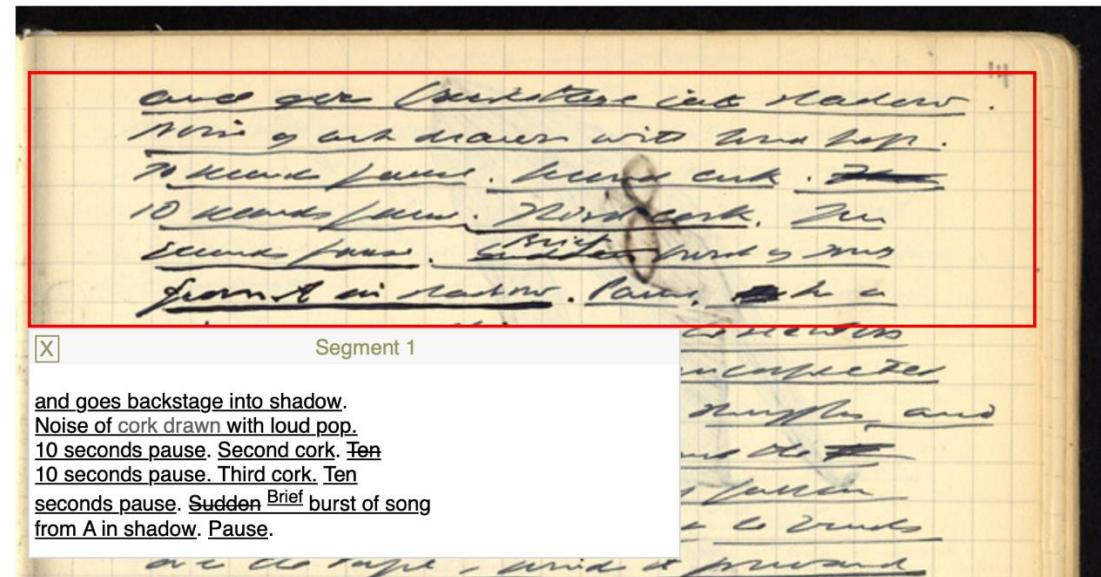


ABOUT ▾ DOCUMENTS ▾ TOOLS ▾ COMPARE SENTENCES

MS-UoR-1227-7-7-1

◀ 14r ▶

[about](#) · [image](#) · [zoom image](#) · [double page view](#) · [image/text](#) · [text](#)



X Segment 1

and goes backstage into shadow.
Noise of cork drawn with loud pop.
10 seconds pause. Second cork. ~~Ten~~
10 seconds pause. Third cork. ~~Ten~~
seconds pause. Sudden ^{Brief} burst of song
from A in shadow. Pause.

unconscious
drifts and
and the ~~is~~
+ tension
- 6 seconds
are do stage - and it proceeds

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

The William Blake Archive is a digital collection of Blake's manuscripts, prints, and other related materials. This screenshot shows a page from 'The Marriage of Heaven and Hell' (Copy G, Printed c. 1818). The left side displays a detailed watercolor illustration of a scene from the manuscript. The right side shows the corresponding diplomatic transcription of the text. The transcription includes line numbers, punctuation, and some annotations. Below the main view, there are links to 'Editor's Notes' and other related objects.

<http://www.blakearchive.org>

ELECTRONIC BEOWULF

The Electronic Beowulf project provides a digital edition of the Old English epic poem. This screenshot shows a page from 'MS 129r, BL 132r'. The left panel shows the manuscript page with the text 'HWÆT WE GAR-DENA IN GEARDAGUM'. The right panel shows the digital transcription of the same text. A glossary window is open at the bottom left, showing definitions for words like 'peodcyniga' and 'þrym gefrunon'. The interface includes navigation tools like 'Goto', 'Edition 129r, 1-21', and 'Fit frame'.

<https://ebeowulf.uky.edu>

TEI for Manuscript Cataloguing



e-codices - Virtual Manuscript Library of Switzerland

The goal of e-codices is to provide free access to all medieval and a selection of modern manuscripts of Switzerland by means of a virtual library.

At the moment, the virtual library contains **2918** manuscripts from **99** different collections.
The virtual library will be continuously updated and extended.

The Virtual Library of Switzerland: e-codices

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

[Introduction](#) - [Technical Help](#) - [Repositories](#) - [Curated Collections](#) - [Search](#)

Help us plan the future of OPenn! We want to learn more about our users, your motivations research projects and goals, and we want to hear from you. Please fill out [this brief survey](#). It should take only 3-5 minutes of your time and will be invaluable to us. — The OPenn Team

[Home](#)

OPenn: Read Me

University of Pennsylvania Libraries: OPenn

[HTTPS://GITHUB.COM/SIMS-MSS/OPENN-XML](https://github.com/sims-mss/openn-xml)

Search Medieval Manuscripts



[ADVANCED SEARCH](#)

OR BROWSE BY

MANUSCRIPTS

WORKS

PEOPLE

PLACES

**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.

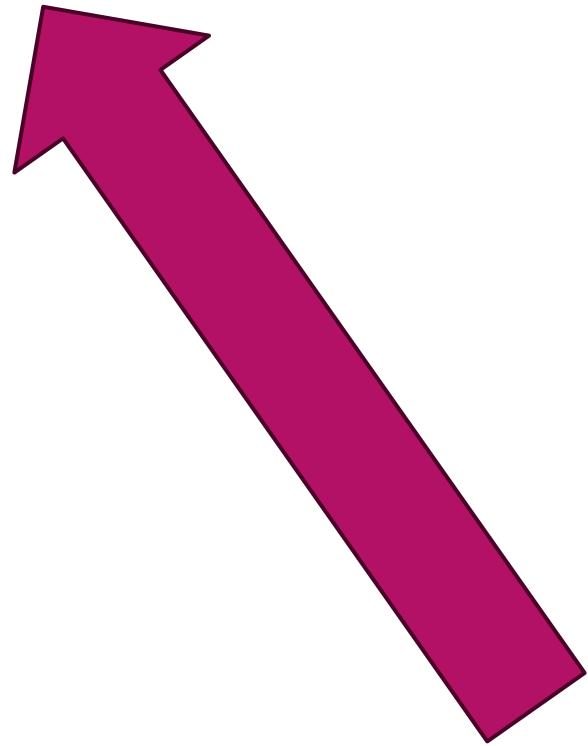
But not everyone

- ▶ 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>
<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>

</textDesc>

<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>

```
<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 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

<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.

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 occurred 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>