

Modelling humanities data with TEI-XML

SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Introductions

► Teacher:

- ▶ Dr Katarzyna Anna Kapitan
- ▶ katarzyna.kapitan [at] chartes.psl.eu

► Students:

- ▶ Name
- ▶ Background and Study Programme
- ▶ Research Interests
- ▶ Experience with Markup Languages (which ones, what experience)

Course overview

► Classes:

- ▶ 10 x 2 hours between 2 October and 11 December
- ▶ Wednesdays, 15:00-17:00
- ▶ @ Campus Condorcet, *Le bâtiment de recherche nord*, Edith Thomas Room (ground floor)

Course overview

- ▶ Course prerequisites:
 - ▶ Laptop with installed
 - ▶ **Oxygen XML Editor**
[\(https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html\)](https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html)
 - ▶ **Sublime Text**
[\(https://www.sublimetext.com/download\)](https://www.sublimetext.com/download)

Objectives

- ▶ Understand the XML structure and its customisation
- ▶ Use the TEI Guidelines for text editing and manuscript cataloguing
- ▶ Create XML templates and validation rules
- ▶ Document encoding choices
- ▶ Apply XML to individual projects

Schedule

Session 1: 2/10	Markup Languages and Text Encoding
Session 2: 9/10	Extensible Markup Language (XML) and Document Type Definition (DTD)
Session 3: 16/10	Text Encoding Initiative (TEI) Guidelines
Session 4: 23/10	Describing Primary Sources
Session 5: 30/10	Transcribing Primary Sources
Session 6: 6/11	Editing Primary Sources
Session 7: 13/11	Annotating Primary Sources
Session 8: 20/11	XPath
27/11	PSL week – no class
Session 9: 4/12	Customisation and Documentation
Session 10: 11/12	Customisation and Validation

Assessment method

► Assessment:

- Attendance (20%)
- Mid-term Assignment (30%) – 30 October
- Final Assignment (50%) – 18 December

Course Materials

► **GitHub Repo:**

https://github.com/KAKDH/ENC_TNAH_2024/

Text Encoding & Markup Languages

Markup Languages

► What is a markup language?

- ▶ Computer language
- ▶ Uses tags to define elements within a document
- ▶ Human-readable

► Examples:

- ▶ HTML (Hypertext Markup Language)
- ▶ Markdown
- ▶ TeX & LaTeX
- ▶ **XML (Extensible Markup Language)**
- ▶ Scribe, GML (Generalized Markup Language) & SGML (Standard Generalized Markup Language)

Markup

	Semantic	Presentational
LaTeX	\selectlanguage{latin}{ad hoc}	\emph{ad hoc}
HTML	< <i>i</i> lang="la">ad hoc</ <i>i</i> >	< <i>i</i> >ad hoc</ <i>i</i> >
XML	<foreign xml:lang="la">ad hoc</foreign>	<hi rend="i">ad hoc</hi>

Exercises

- ▶ Download the Exercises Folder from:
<https://tinyurl.com/Wk1Ex>

Exercise Week1_Ex1

Individual Work

(<https://tinyurl.com/Wk1Ex>)

- ▶ Which of these files contain markup?
- ▶ What type of markup is used?
- ▶ How is the markup expressed?
- ▶ How is the document structured?
- ▶ What are these types of markup usually used for, give examples?
- ▶ List all markup elements used in the documents and explain what are they used for (what do you think they mean)?

Example 1:
AM_30_fol.txt
AM_30_fol.md
AM_30_fol.html
AM_30_fol.doc

Example 2:
AM_30_fol.txt
AM_30_fol.md
AM_30_fol.tex
AM_30_fol.doc

Group work

- ▶ Create two groups. One group consisting of everyone who worked on example 1 and one group consisting of everyone who worked on example 2.
- ▶ Compare your answers and prepare a presentation of your example, choose one person who will present your answers.
- ▶ **Group 1:** focus on HTML
- ▶ **Group 2:** focus on TEX

Plain Text | Formatted Text

- ▶ **Plain text** format contains no formatting information,
- ▶ **Formatted Text - Rich text** format includes formatting details such as font size, style, colour, etc.
- ▶ **WYSIWYG (What You See Is What You Get)** - a formatted document as it will appear on screen or in print without showing the descriptive code.

Typesetting with TeX & LaTeX

- ▶ TeX is a typesetting system widely used in academia, especially in mathematics, computer science, engineering, linguistics, and many more, among them: **digital scholarly editing**.
- ▶ **WYSIWYM** (What You See Is What You Mean) is a paradigm for editing a structured document separating presentation from content.
- ▶ TeX commands commonly start with a backslash and are grouped with curly braces.
 - ▶ **\myCommand{ My Content goes here }**

Saga af Hrómu(n)de Greipszýne

Cäpituli .I.

(S)á[¶] kongr rieþe fyrir |Gordom| ¶|l| ¶|danmorc| ¶ er Olaf hiet,[¶]
hann var sonr Gnófar Asmundar, hann var frøgt *mahr*.[¶] Brøðr
.ij. kári oc ørnulfr, voru landvarnar menn kongz, hermenn mikler.
5 Par Bió eirn rýkur Boandi, sá hiet Greýpr. hann átte þá *kono*, er
Gunnloþ hiet, dótter Hrókz hinz suarta, Þau átto .ix. soho er sva
hieto. Hrólf, Hake, Gautr, Þróstr, Angantýr, Logi, Hrómundr.
Helge. Hrókr. Þeir voru aller efnileger menn. Þo var Hrómundr
fyrir þeim aullom. hann kunni eigi at hrödast, hann var augna fagr,
10 hárbíarr, oc herþamikill, mikill oc stercr, lýktiz mið Hróki móþr
fauþr sýnom. Med kongi voru .ij. menn, hiet eirn Býldur, annar
voli, Þeir voru Iller oc underføruler. kongr matti Þá mikils. Eyyt
sinn hiellt olafr kongr, austur[¶] fyrir noreg med her sinn, oc hielldo
aþ Vlfaskerium, herioþo, oc lau viþ eitt Eýland. Kongr býþr Kára
15 oc Ørnulfi aþ ganga uppá Eýuna, oc vita, huort þeir sæ einginn
herskip. Þeir gengu uppá *landiþ*, oc litu .vj. herskip under hómrum
nocrum. Þar var eirn dreke allskrautligr. Kári kallar til þeira, oc spir
huorier fyrir skiponom rieþi, Eirn dólgr stóþ uppá drekanom, oc
quadz Hraungviþr heita. eþr huort er nafn þitt. Kari sagpi til sýn
20 oc sýnz broþurz. oc mælti. Eg veit aungvann verri enn þik, oc þar
fyrir skal ek hauggya þic i smá sticki, Hraungviþr. mælti: Ek hefi

Kapittel 10. Modellering av myndleidir AML

Kapitän, Modelling humanities data with TEI-XML

1



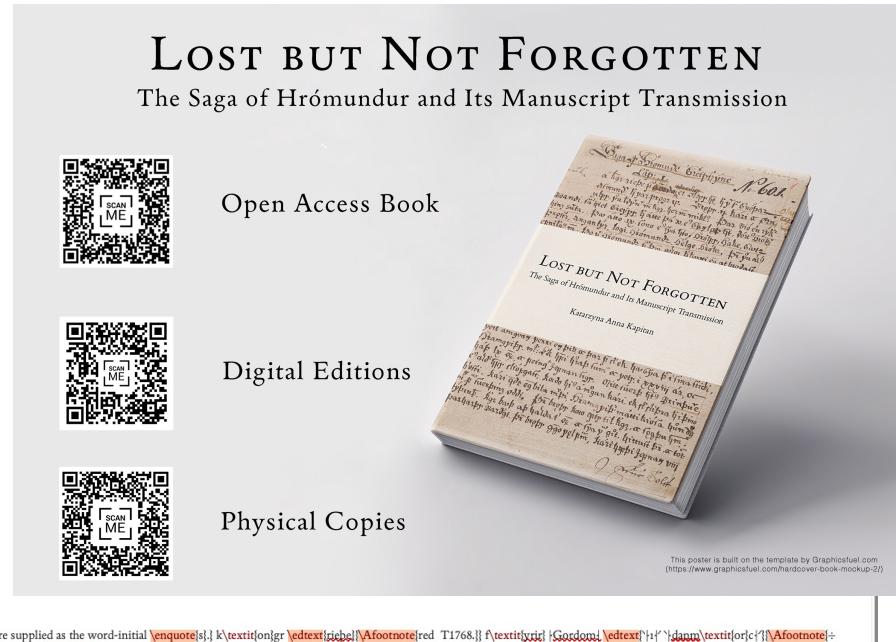
Open Access Book



Digital Edition

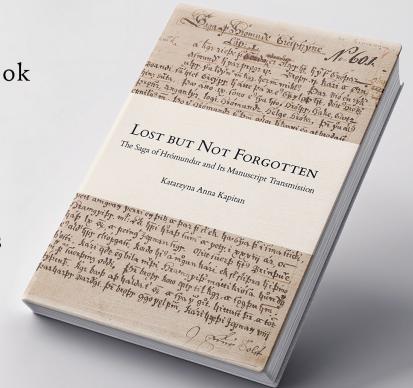


Physical Copies



LOST BUT NOT FORGOTTEN

The Saga of Hrómundur and Its Manuscript Transmission



This poster is built on the template by Graphicsfuel.com
(<https://www.graphicsfuel.com/hardcover-book-mockup-2/>)

HTML (Hypertext Markup Language)

- ▶ HTML is the standard markup language for web pages.
- ▶ Markup:
 - ▶ **<tagname> Content goes here... </tagname>**
 - ▶ Start Tag --- Content --- End Tag



```
<!DOCTYPE html>
<html>
  <head>
    <title> This is my title</title>
  </head>
  <body>
    <p>This is a paragraph</p>
  </body>
</html>
```

Basic elements:

- <html> – root element of an HTML page
- <head> – contains meta information about the document
- <body> – contains the visible page content

Other elements (examples):

- <div> - division/section of the page
- <table> - table
- - image
- – list
- <h1> – heading
- <p> – paragraph

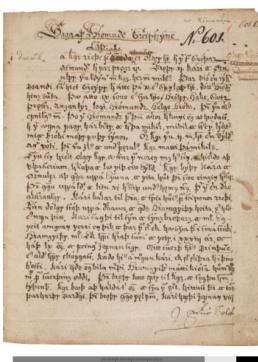
HTML is everywhere !!!

TAYLOR EDITIONS Home Editions Topics Training About

Hrómundar saga Greipssonar, based on AM 601 b 4to with variants

Facsimile Transcription Translation About Downloads

Facsimile (A601)



https://editions.mml.ox.ac.uk/editions/hromundar_A601/

Kapitan, Modelling humanities data with TEI-XML

Clear separation of content and presentation

```
<div>
```

```
    <h2>Short Description</h2>
```

```
        <p>AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, which consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark.
```

```
    </p>
```

```
</div>
```

Clear separation of content and presentation

HTML + CSS (Cascading Style Sheets)

► HTML

```
<h1> My Text </h1>
```

► CSS

```
h1 {  
    text-decoration: underline;  
    text-align: center;  
}
```

Exercise Week1_Ex2

(<https://tinyurl.com/Wk1Ex>)

- ▶ Open AM_30_fol.html with your browser – take a screenshot of the view
- ▶ Open AM_30_fol.html with Sublime Text
- ▶ Associate style.css with your html file by adding within <head> the element <link>:
`<link rel="stylesheet" type="text/css" href="style.css">`
- ▶ Save the file.
- ▶ Open AM_30_fol.html in a browser – take a screenshot of the view.

```

1 1<!DOCTYPE html>
2 <html>
3 <head>
4   <meta charset="utf-8"/>
5   <meta name="author" content="Katarzyna Anna Kapitan">
6   <title>Copenhagen, Den Arnamagnæanske Samling, AM 30 fol.</title>
7 </head>
8 <body>
9   <h1>Copenhagen, Den Arnamagnæanske Samling, AM 30 fol.</h1>
10 <div>
11 <h2>Short Description</h2>
12 <p>AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, with consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark. AM 30 fol. contains two main texts dealing with the Christianization of the Slavs.</p>
13 </div>
14 <div>
15 <h2>Contents</h2>
16 <p>
17 <ul>
18   <li><i>Chronica Slavorum</i> b</li>
19   <li><i>Chronica Slavorum</i> b</li>
20 </ul>
21 </p>
22 </div>
23 </body>
24 </html>

```



Copenhagen, Den Arnamagnæanske Samling, AM 30 fol.

Short Description

AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, with consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark. AM 30 fol. contains two main texts dealing with the Christianization of the Slavs.

Contents

- *Chronica Slavorum* by Helmold of Bosau
- *Chronica Slavorum* by Arnold of Lübeck

COPENHAGEN, DEN ARNAMAGNÆANSKE SAMLING, AM 30 FOL.

SHORT DESCRIPTION

AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, with consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark. AM 30 fol. contains two main texts dealing with the Christianization of the Slavs.

CONTENTS

- *Chronica Slavorum* by Helmold of Bosau
- *Chronica Slavorum* by Arnold of Lübeck

HTML is everywhere: Behind the e-codices viewer

The screenshot shows the e-codices viewer interface. At the top, there's a search bar with the query "Krakow, Jagiellonian Library, Depositum (Ms. Berol. Theol. Lat. Qu. 1)". Below the search bar, there are tabs for "Document Details", "Annotations", and "Additional Bibliography". On the left, there's a thumbnail of a manuscript page showing a circular emblem with Latin text. To the right of the thumbnail, there's a sidebar with fields like "Country of Location: Poland", "Location: Krakow", and "Manuscript Summary: ...". Below the sidebar, there's a detailed description of the manuscript, including its title, date of origin, and dimensions.

This screenshot shows the e-codices viewer with the TEI XML source code visible. The XML is a representation of the data shown in the viewer. It includes sections for the title, date, and publisher information, as well as specific details about the manuscript's physical appearance and its digital representation.

```
<TEI version="5.1" xsi:schemaLocation="http://www.tei-c.org/ns/1.0 ./xsd/TEI-PS/1.7/tei-p5-e-codices_1.7.xsd" xml:lang="deu" xml:id="eCod_bj-Berol-Theol-Lat-Qu-0001">
  ...
  <fileHeader>
    ...
  </fileHeader>
  <fileDesc>
    ...
  </fileDesc>
  <titleStmt>
    <title>Epistolar</title>
    ...
  </titleStmt>
  <editionStmt>
    <edition>Elektronische Version nach TEI P5.1</edition>
    ...
  </editionStmt>
  <respStmt>
    ...
  </respStmt>
  <resp>
    Konvertierung nach TEI:
    <persName>Michael Wiedkehr</persName>
    <date when="2015-11-26">26.11.2015</date>
  </resp>
  <name>
    e-codices - Virtual Manuscript Library of Switzerland
  </name>
  <respStmt>
    ...
  </respStmt>
  <editionStmt>
    ...
  </editionStmt>
  <publicationStmt>
    ...
  </publicationStmt>
  <publisher>
    e-codices - Virtual Manuscript Library of Switzerland
  </publisher>
  <availability status="restricted" n="cc-by">
    <licence target="http://creativecommons.org/licenses/by/3.0/">
      ...
    </licence>
  </availability>
  <sourceDesc>
    ...
  </sourceDesc>
  <bibl>PDF vorhanden</bibl>
  <msDesc xml:id="bj-Berol-Theol-Lat-Qu-0001_vonEuw" xml:lang="deu">
    ...
  </msDesc>
  <msIdentifier>
    ...
  </msIdentifier>
  <settlement>Krakow</settlement>
  ...
</TEI>
```

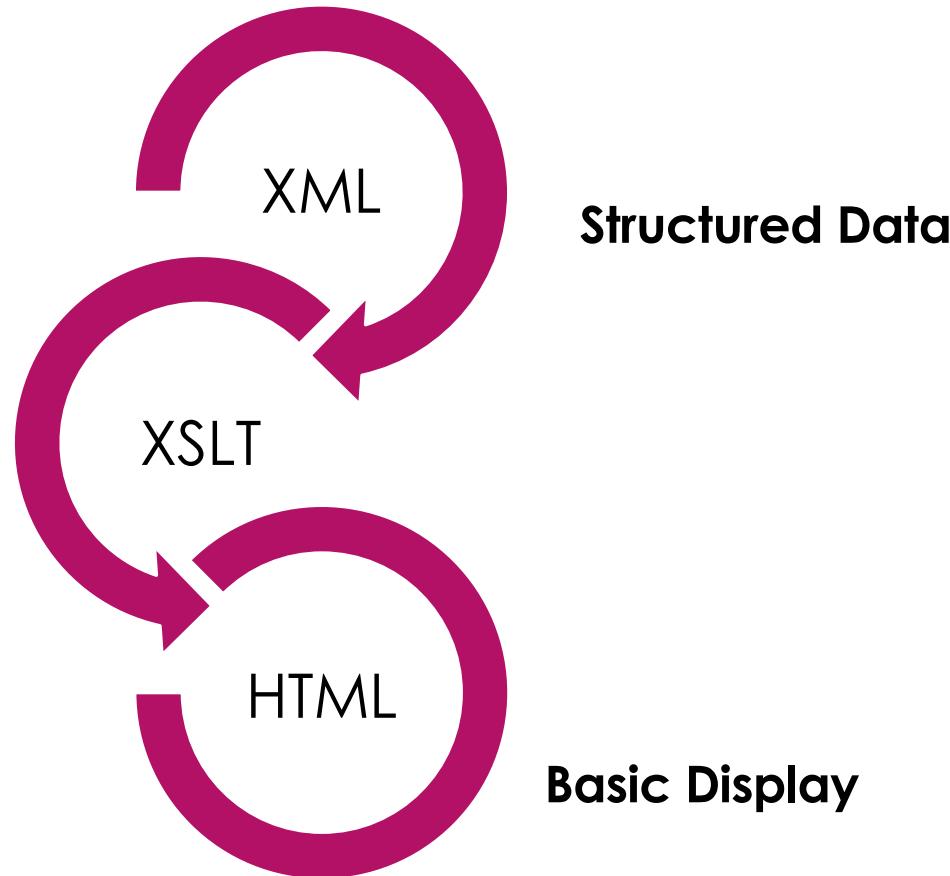
Kraków, Jagiellonian Library, Ms. Berol. Theol. Lat. Qu. 1

Source: <https://www.e-codices.unifr.ch/en/list/one/bj/Berol-Theol-Lat-Qu-0001/>
Kapitan, Modelling humanities data with TEI-XML

From Data to Display



Transformation Scenarios



Structured Data

Basic Display

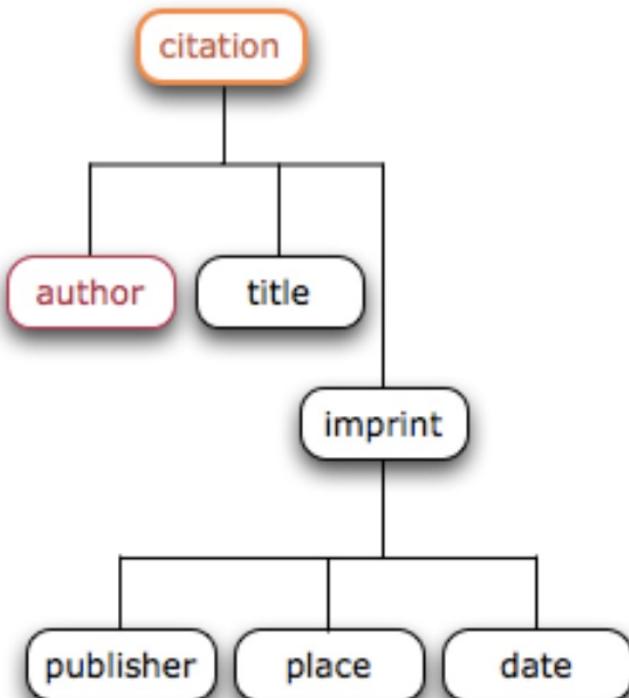
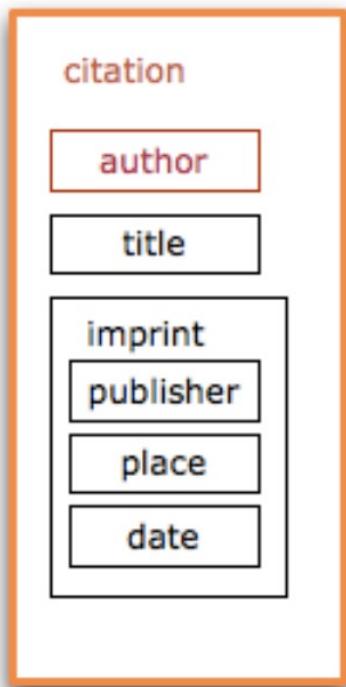
XML **(Extensible Markup Language)**

XML (Extensible Markup Language)

- ▶ Storing structured data
- ▶ International standard, non-proprietary
- ▶ Standard text format (expressed in plain text)
- ▶ Easy to parse and read for computer programs.
- ▶ Widely used to export and share structured data.
- ▶ Hardware and software independent



Syd Bauman, Julia Flanders, and WWP, Creative Commons Attribution-ShareAlike 3.0 (Unported) license.



```

<?xml version="1.0" encoding="UTF-8"?>
<citation>
  <author>Katherine Hayles</author>
  <title>Writing Machines</title>
  <imprint>
    <publisher>MIT Press</publisher>
    <place>Cambridge, MA</place>
    <date>2002</date>
  </imprint>
</citation>
  
```

Syd Bauman, Julia Flanders, and WWP, Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

XML Elements & Tags

```
<!DOCTYPE html>
```

```
<html>
```

```
  <head>
```

```
    <title>This is a title</title>
```

```
  </head>
```

```
<body>
```

```
  <p>This is a paragraph</p>
```

```
</body>
```

```
</html>
```

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

```
<myRoot>
```

```
  <myContent>
```

```
    <content>
```

```
      Here is my content
```

```
    </content>
```

```
  </myContent>
```

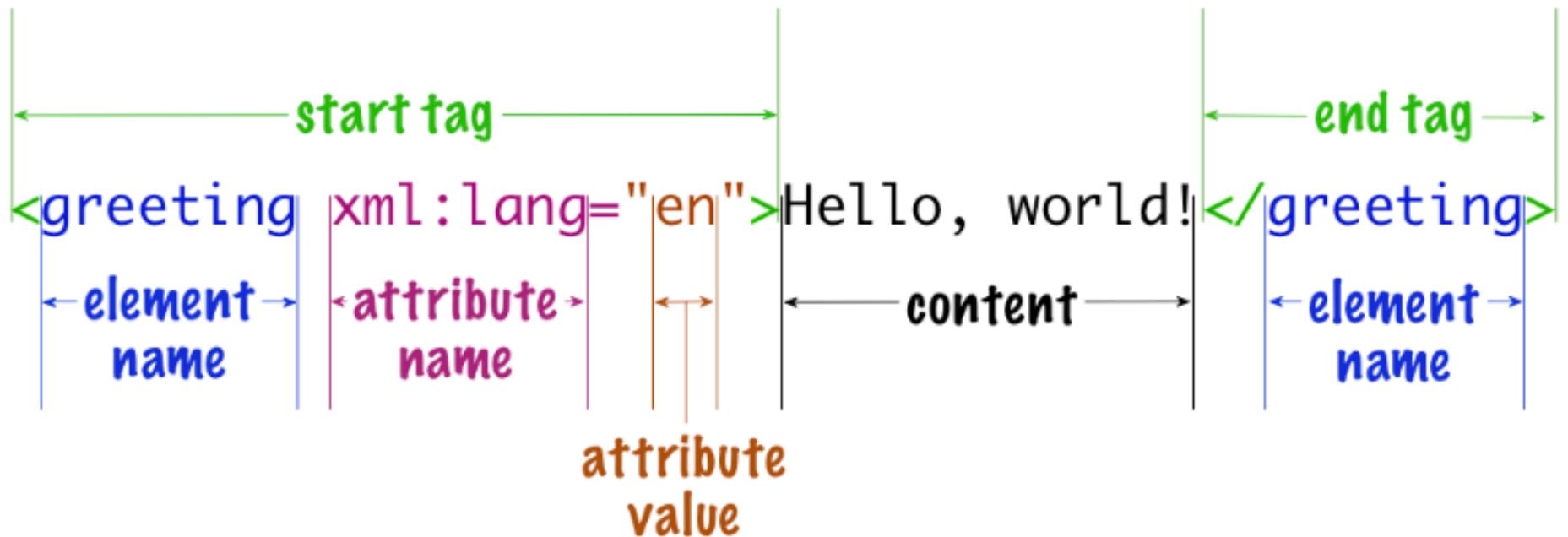
```
</myRoot>
```

XML Elements & Tags

- ▶ Text is divided into elements (the nouns of the encoding — content objects).
- ▶ elements have **start-tags** and **end-tags**
 - ▶ `<heading>My heading</heading>`
- ▶ start-tags have `< ... >`
 - ▶ `<heading>`
- ▶ end-tags have `</ ... >`
 - ▶ `</heading>`

XML Attributes

- ▶ Attributes are adjectives of XML, they describe the properties of the elements
 - ▶ any number of attributes can be specified on a given start-tag
 - ▶ but only one with a given name
-
- ▶ **<person job="musician" age="55">Paul Simon</person>**



Syd Bauman, Julia Flanders, and WWP, Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

Quiz

I. Which of the following examples are well-formed XML?

1. <name type="person">Pearl S. Buck</name>
2. <name type="person">Toni Morrison</name>
3. <name="person">Carl Sagan</name>
4. <name type="person">Kurt Vonnegut</name>
5. <name type=person>John Cleese</name>
6. <name type="person"><forename>Frances</forename><surname>Perkins</surname></name>

II. List the element names, attribute names, attribute values, contents

Clear Separation of Content and Presentation

```
<?xml version="1.0" encoding="UTF-8"?>  
<myRoot>  
    <myContent>  
        <content>  
            Here is my content  
        </content>  
    </myContent>  
</myRoot>
```

```
<?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>
      <affiliation/>
    </name>
    <name>
      <firstName>Anna</firstName>
      <lastName>Smith</lastName>
      <affiliation/>
    </name>
    <name>
      <firstName>Jan</firstName>
      <lastName>Kowalski</lastName>
      <affiliation/>
    </name>
  </participants>
</workshop>
```

Kapitan, Modelling humanities data with TEI-XML

XML Structure

Workshop:
Instructors:
KAK
Participants:
JD
AS
JK



```
<msDesc>
```

```
  <msIdentifier>
```

```
    <settlement>Copenhagen</settlement>
```

```
    <repository>Den Arnamagnæanske Samling</repository>
```

```
    <idno>AM 30 fol.</idno>
```

```
  </msIdentifier>
```

```
  <msContents>
```

```
    [...]
```

```
    <msItem>
```

```
      <title>Chronica Slavorum</title>
```

```
      <author>Helmold of Bosau</author>
```

```
    </msItem>
```

```
    <msItem>
```

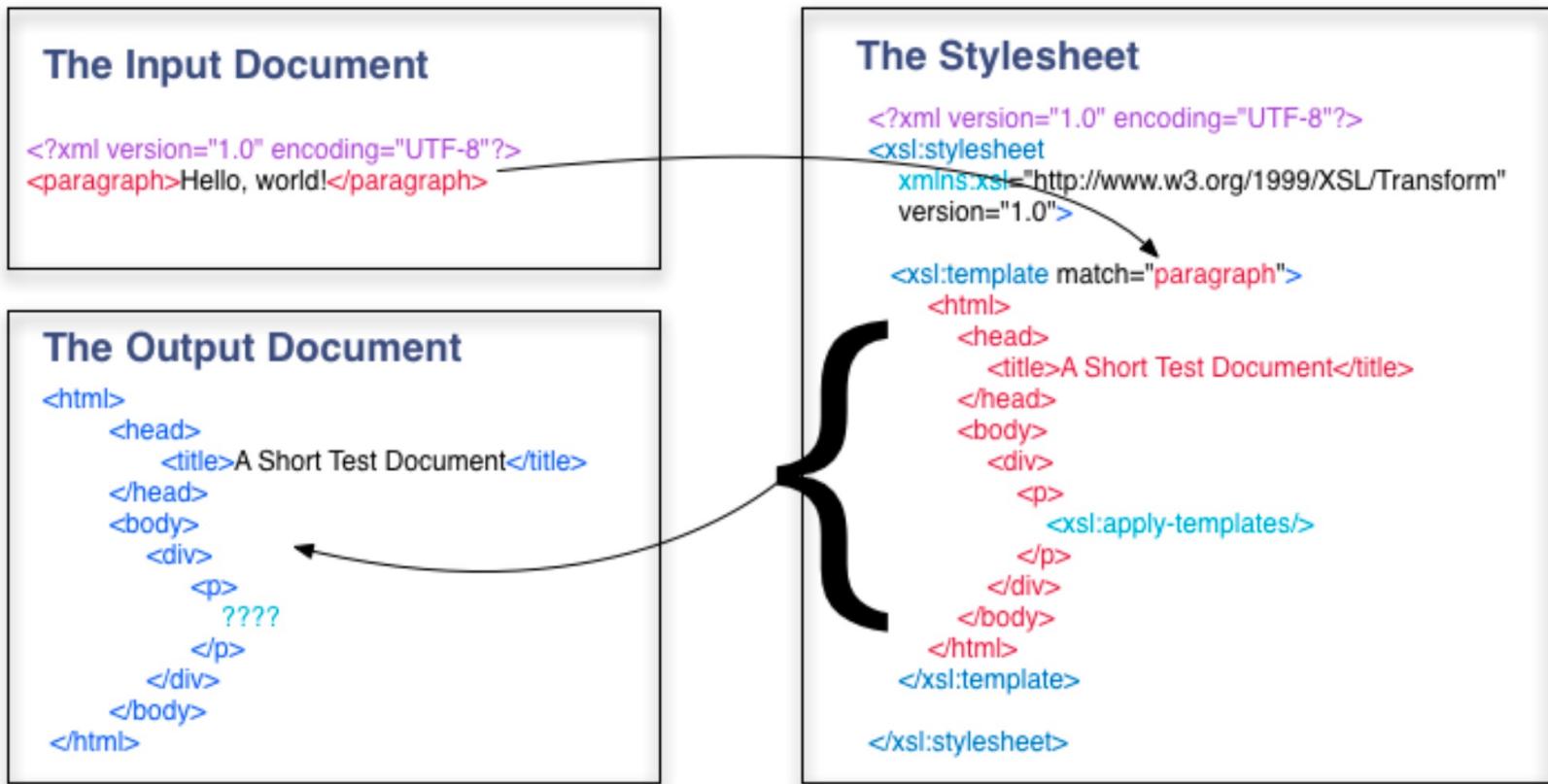
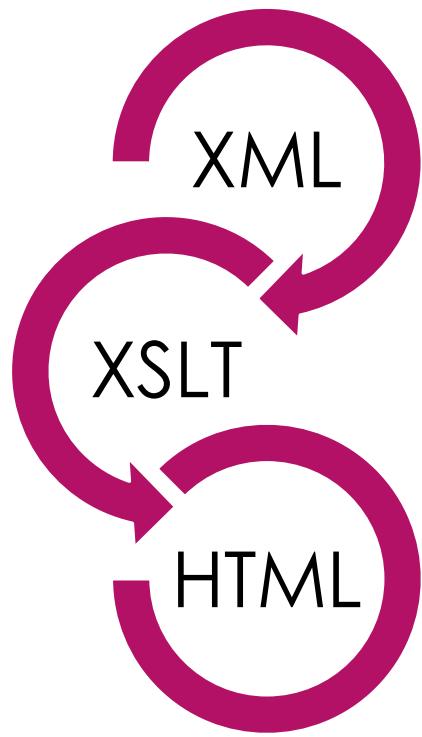
```
      <title>Chronica Slavorum</title>
```

```
      <author>Arnold of Lübeck</author>
```

```
    </msItem>
```

```
  </msContents>
```

```
</msDesc>
```

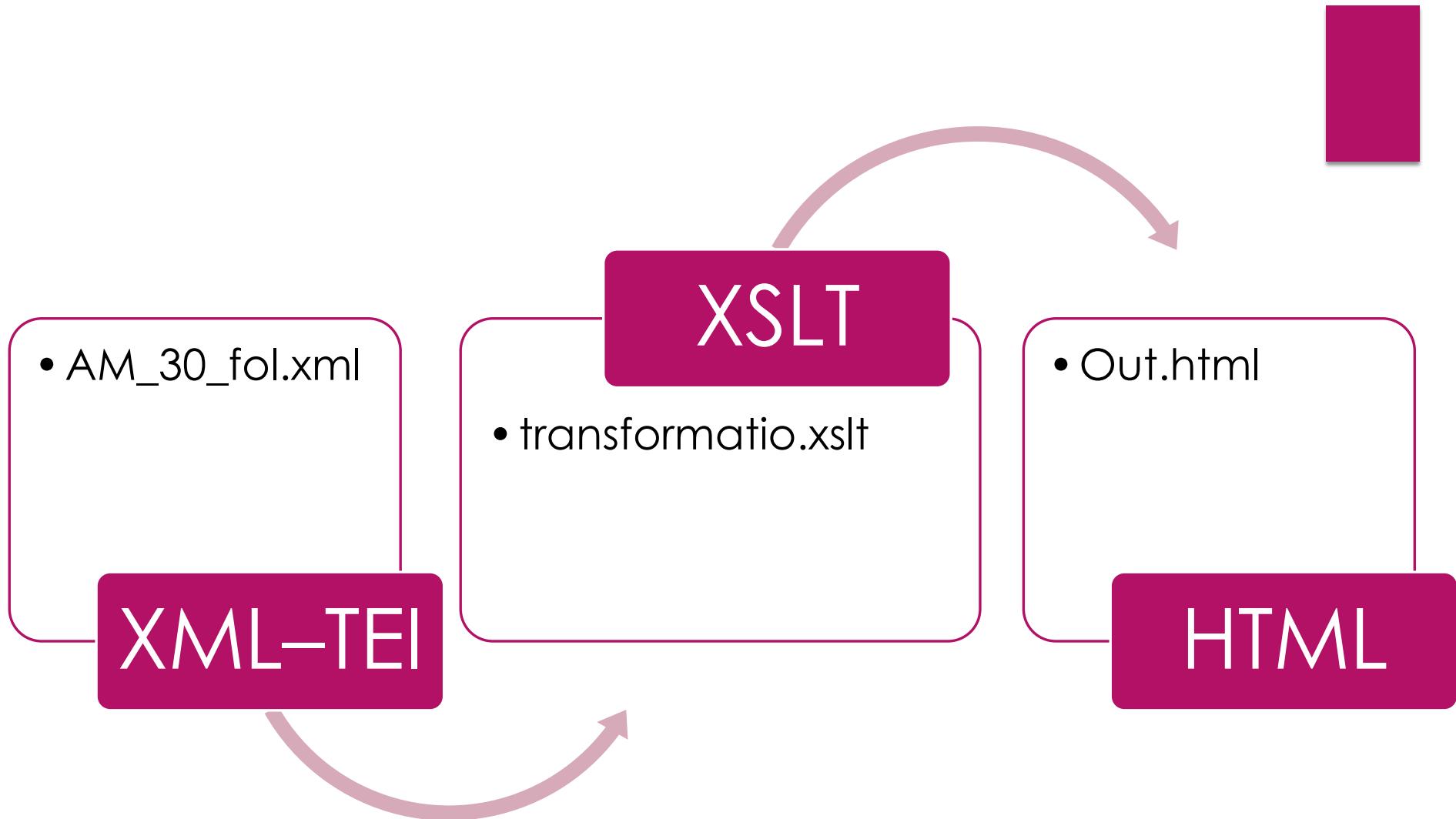


© 2012 Syd Bauman, Julia Flanders, and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

Exercise Week1_Ex3

(<https://tinyurl.com/Wk1Ex>)

- ▶ Convert an XML file to HTML with XSLT
- ▶ Follow the guidelines in:
“Guidelines_Transformation_Scenario.pdf”



COPENHAGEN, DEN ARNAMAGNEANSKE SAMLING, AM 30 FOL.

SHORT DESCRIPTION

AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, which consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark. AM 30 fol. contains two main texts dealing with the Christianization of the Slavs.

CONTENTS

- *Chronica Slavorum by Helmold of Bosau*
- *Chronica Slavorum by Arnold of Lübeck*



Our initial HTML file:
AM_30_fol.html

AM_30_fol.xml

```

  TEI  teiHeader  fileDesc  titleStmt
  19   <idno>AM 30 fol.</idno>
  20   </msIdentifier>
  21 <msContents>
  22   <summary>AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, which consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark. AM 30 fol. contains two main texts dealing with the Christianization of the Slavs.</summary>
  23   <msItem>
  24     <title>Chronica Slavorum</title>
  25     <author>Helmold of Bosau</author>
  26   </msItem>
  27   <msItem>
  28     <title>Chronica Slavorum</title>
  29
  30
  31
  32 </msItem>
  33

```

Text Grid Author

Our XML file:
AM_30_fol.xml

COPENHAGEN, DEN ARNAMAGNEANSKE SAMLING, AM 30 FOL.

SHORT DESCRIPTION

AM 30 fol. is a paper manuscript in folio format comprising 56 leaves gathered into five quires of six conjoint leaves each, with the exception of the last quire, which consists of four conjoint leaves only. The whole codex is made of one type of relatively thick laid paper with a "bull's head" watermark. AM 30 fol. contains two main texts dealing with the Christianization of the Slavs.

CONTENTS:

- *Chronica Slavorum by Helmold of Bosau*
- *Chronica Slavorum by Arnold of Lübeck*



An HTML file - Result of the transformation of our XML file with XSLT: out.html



Course Materials:

https://github.com/KAKDH/ENC_TNAH_2024/

Assignments:

https://github.com/KAKDH/ENC_TNAH_2024/blob/main/Documents/ENC_Kapitan_TEI-XML_Assignments.pdf

Schedule and bibliography:

https://github.com/KAKDH/ENC_TNAH_2024/blob/main/Documents/ENC_Kapitan_TEI-XML.pdf