Modelling humanities data with TEI-XML

SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Dr Katarzyna Anna Kapitan
2 October 2024

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)
    - **Sublime Text** (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	*ad hoc*	*ad hoc*
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** (**W**hat **Y**ou **S**ee **I**s **W**hat **Y**ou **G**et) - 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

Cäp*ituli* .I.

1r

⟨S⟩á[1] kongr rieþe *fyrir* ⸠Gordom⸡ ᛘᛁᛚ ᛘdanmorcᛚ[2] er Olafr hi*et*,[3] h*ann* var *sonr* Gnóþar Asmund*ar*, h*ann* var frǫgr *maþr*.[4] Brǫþr .ij. kári oc ørnulfr, v*oru* l*and*var*nar* m*enn* kongz, herm*enn* mikl*er*. Þar Bió ei*rn* rýk*ur* Boandi, sá hiet Greýpr. h*ann* átte þá *kono*, e*r* Gunnlǫþ hi*et*, dótt*er* Hrókz hinz su*ar*ta, Þau átto .ix. sono e*r* sva hi*et*o. Hrólfr, Hake, Gautr⟨,⟩ Þrǫstr, Angantýr, Logi, Hrómundr. Helge. Hrókr. Þ*eir* v*oru* all*er* efnil*eger* m*enn*. Þo v*ar* Hrómundr *fyrir* þeim aullom. h*ann* kunni e*igi* at hrǫdast, h*ann* v*ar* augna fagr, hárbi*ar*tr, oc h*er*þamikill, mikill oc st*er*cr, lýkt*iz* miǫc Hróki móþr fauþr sýnom. *Med* kongi v*oru* .ij. m*enn*, hi*et* ei*rn* Býld*ur*, ann*ar* voli, Þ*eir* v*oru* Ill*er* oc und*er*fǫrul*er*. kongr matti Þá mikils. Eytt sinn hiellt olafr kongr, austur[5] *fyrir* noreg m*ed* h*er* sinn, oc hielldo aþ Vlfaskerium, h*er*ioþo, oc lau viþ eitt Eýl*and*. Kongr býþr Kára oc Ørnulfi aþ g*an*ga uppá Eýuna, oc vita, hu*or*t þ*eir* sæe einginn h*er*skip. Þ*eir* gengu uppá l*and*i*þ*, oc litu .vj. h*er*skip und*er* hǫm*rum* n*ocrum*. Þ*ar* v*ar* ei*rn* dr*e*ke allskrautl*igr*. Kári kallar til þ*eir*a, oc spir hu*orier* *fyrir* skiponom rieþi, Ei*rn* dólgr stóþ uppá dr*e*kanom, oc q*va*dz Hraungviþr heita. e*þr* hu*or*t e*r* nafn þitt. Kari sagþi til sýn oc sýnz broþurz. oc mæl*ti*. Eg veit aungvann verri enn þik, oc þar *fyrir* *ska*l ek hauggva þ*ic* i smá sticki, Hraungviþr. mæl*ti*: Ek hefi

---

2 rieþe] red T1768. 2 ᛘᛁᛚ ᛘdanmorcᛚ] ÷ T1768, A345, P67, L222, B4859. 3 Asmund*ar*] + d. A587, A193. 3 frǫgr *maþr*] maþr frægr A587, A193. 5 sá hiet Greýpr] er Greypur hiet A345. 9 eigi] ei A193, P67, L222. 13 kongr] ÷ A345. 13 austur] norþr A345, vestur P67. 15 sæe] sæo, A345, B4859. 15 einginn] eingi A587, engi A193, T1768, A345, L222, B4859. 16 h*er*skip] skip A345, L222, B4859. 17 spir] + efter P67. 18 hu*orier*] hver A587, A193, hvorir T1768, hvor A345, P67, L222, huerier B4859. 21 *ska*l] vil P67. 21 Hraungviþr. … L222, B4859.

Lost but Not Forgotten

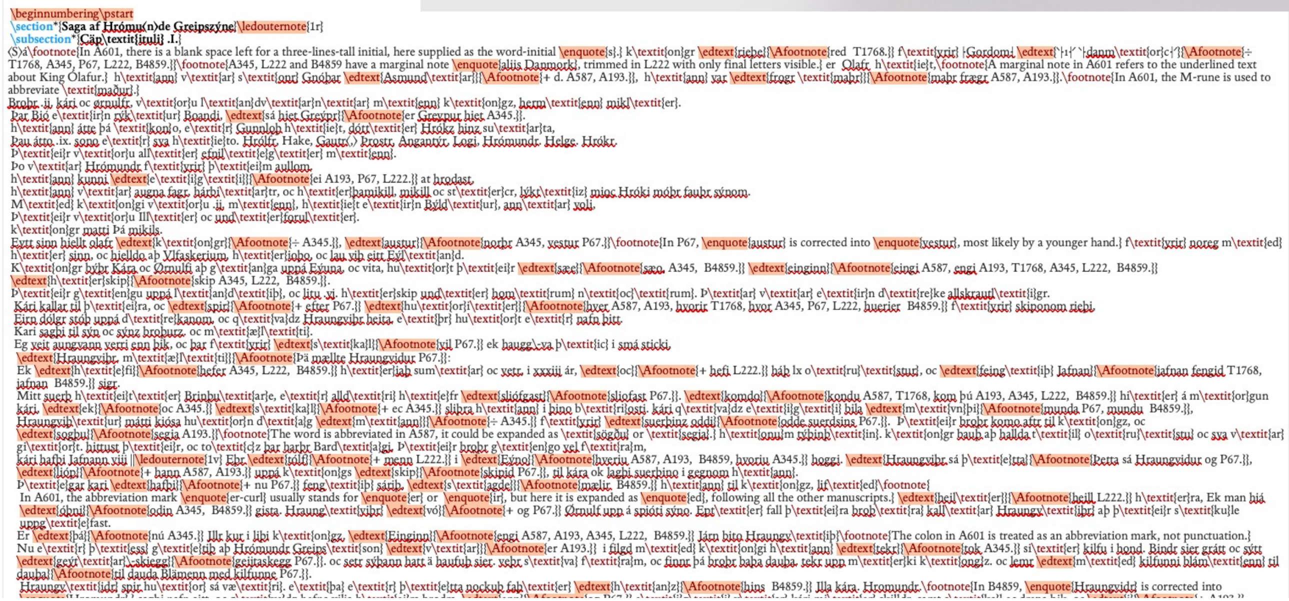
The Saga of Hrómundur and Its Manuscript Transmission

Open Access Book

Digital Editions

Physical Copies

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>
```

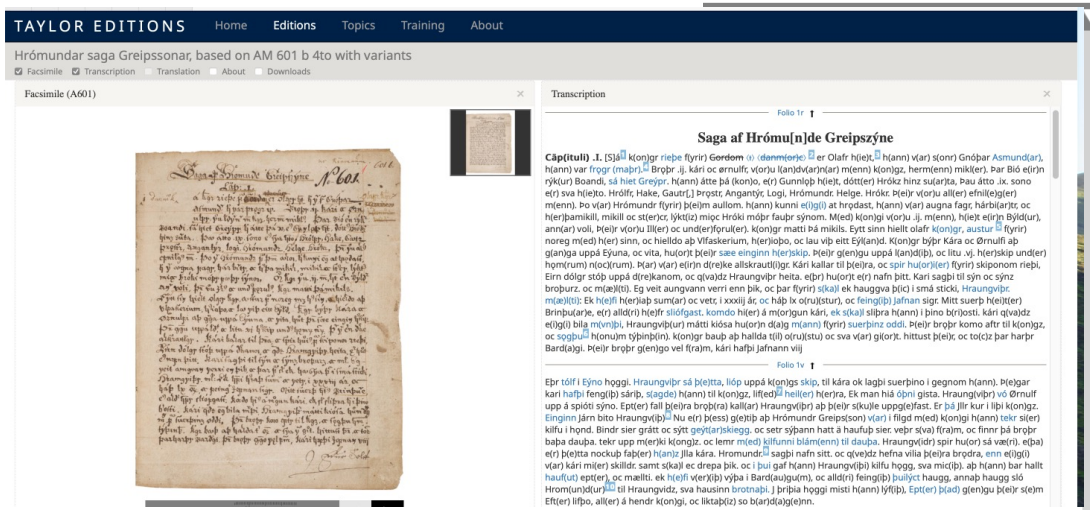
Basic elements:

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

Other elements (examples):

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

HTML is everywhere !!!




```
▼ <div class="row" style="height:inherit;">
    ::before
  ▼ <div class="col-xs-12">
    ▼ <div id="text" class="panel-group" role="tablist">
      ▶ <div id="facsimile" class="panel panel-default col_rs ui-widget ui-widget-content
      ui-helper-clearfix ui-corner-all">…</div>
      ▼ <div id="transcription" class="panel panel-default col_rs ui-widget ui-widget-
      content ui-helper-clearfix ui-corner-all">
        ▶ <div class="panel-heading ui-sortable-handle ui-widget-header ui-corner-all"
        role="tab">…</div>
        ▼ <div class="panel-body" role="tabpanel"> Scroll Event
            ::before
          ▼ <div class="tei_body">
            ▶ <span class="pagebreak" id="index.xml-pb-d65476e258">…</span> flex
            ▶ <h1 class>…</h1>
            ▼ <section class="chapter" id="index.xml-body.1_div.1">
```

```
▼ <p> = $0
  ▼ <span class="rubric" itemprop="hi">
      "Cäp"
      <span class="ex" itemprop="ex">(ituli)</span>
      " .I."
    </span>
    <span class="supplied" itemprop="supplied">[S]</span>
    "á"
  ▼ <span id="ftn001_return">
    ▶ <a class="notelink collapsed" title="In A601, there is a blank space
    left for a three-lines-tall initial, here supplied as the word-initial
    s." href="#ftn001" aria-controls="ftn001" aria-expanded="false" data-
    toggle="collapse">…</a> Event
    </span>
  ▼ <span class="collapse note " id="ftn001">
      <sup>1</sup>
      " In A601, there is a blank space left for a three-lines-tall initial,
      here supplied as the word-initial "
      <span class="q" itemprop="q">s</span>
      "."
    </span>
    " k"
    <span class="ex" itemprop="ex">(on)</span>
```

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

Clear separation of content and presentation

<h2>Short Description</h2>

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.

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.

```
<!DOCTYPE html>
<html>
<head>
   <meta charset="utf-8"/>
   <meta  name="author" content="Katarzyna Anna Kapitan">
<title>Copenhagen, Den Arnamagnæanske Samling, AM 30 fol.</title>
</head>
<body>
      <h1>Copenhagen, Den Arnamagnæanske Samling, AM 30 fol.</h1>
<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, 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>
</div>
<div>
   <h2>Contents</h2>
<p>
<ul>
   <li><i>Chronica Slavorum</i> b
   <li><i>Chronica Slavorum</i> b
</ul>
</p>
</div>
</body>
</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.

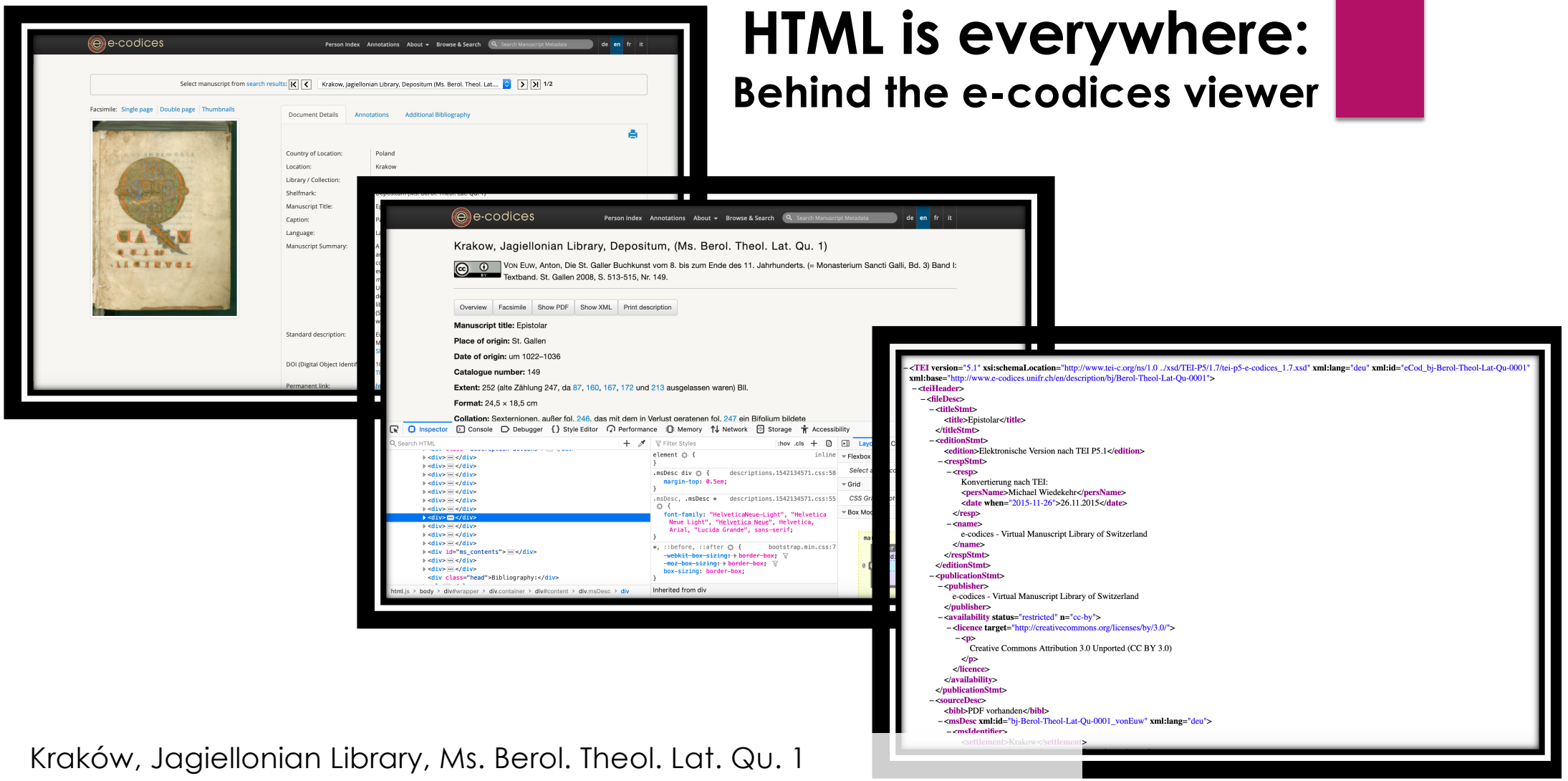
Short Description

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

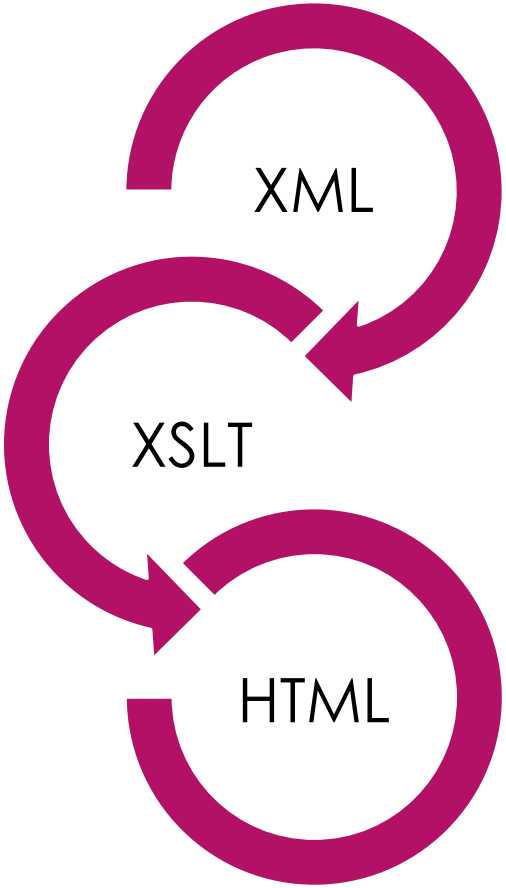
- Chronica Slavorum by Helmold of Bosau
- Chronica Slavorum by Arnold of Lübeck

HTML is everywhere:
Behind the e-codices viewer



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/

From Data to Display



Structured Data

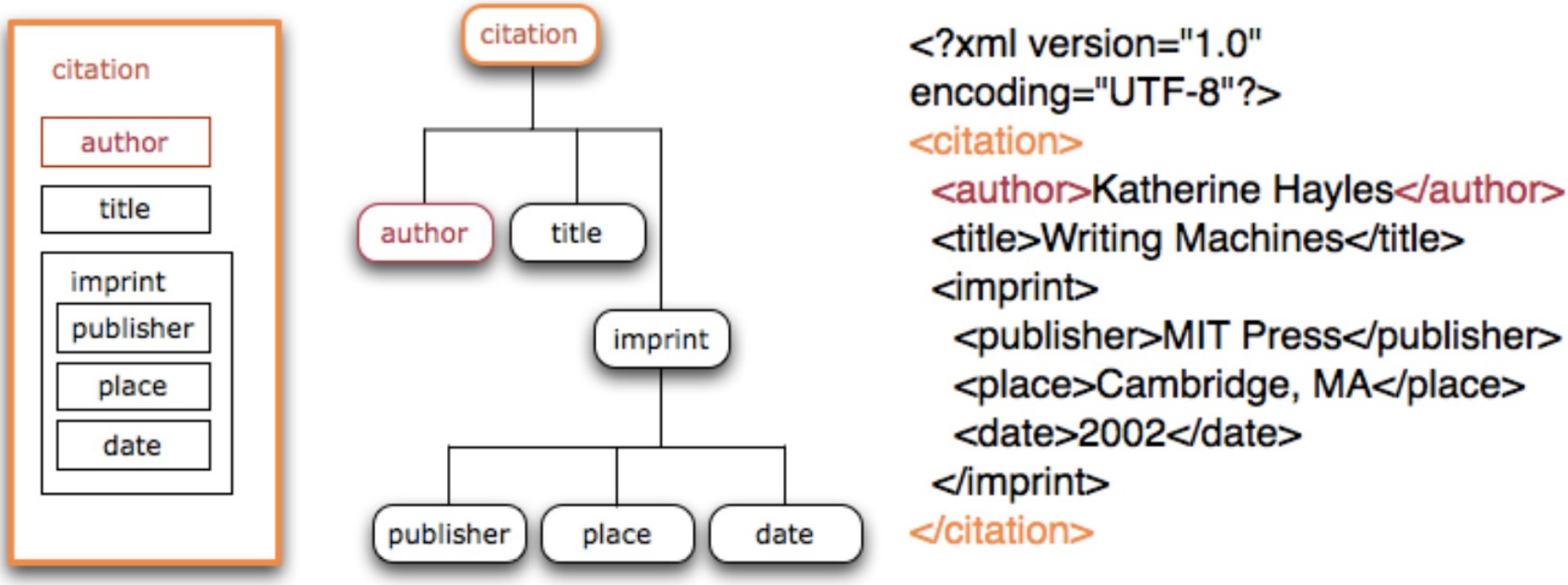
Transformation Scenarios

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



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>
```

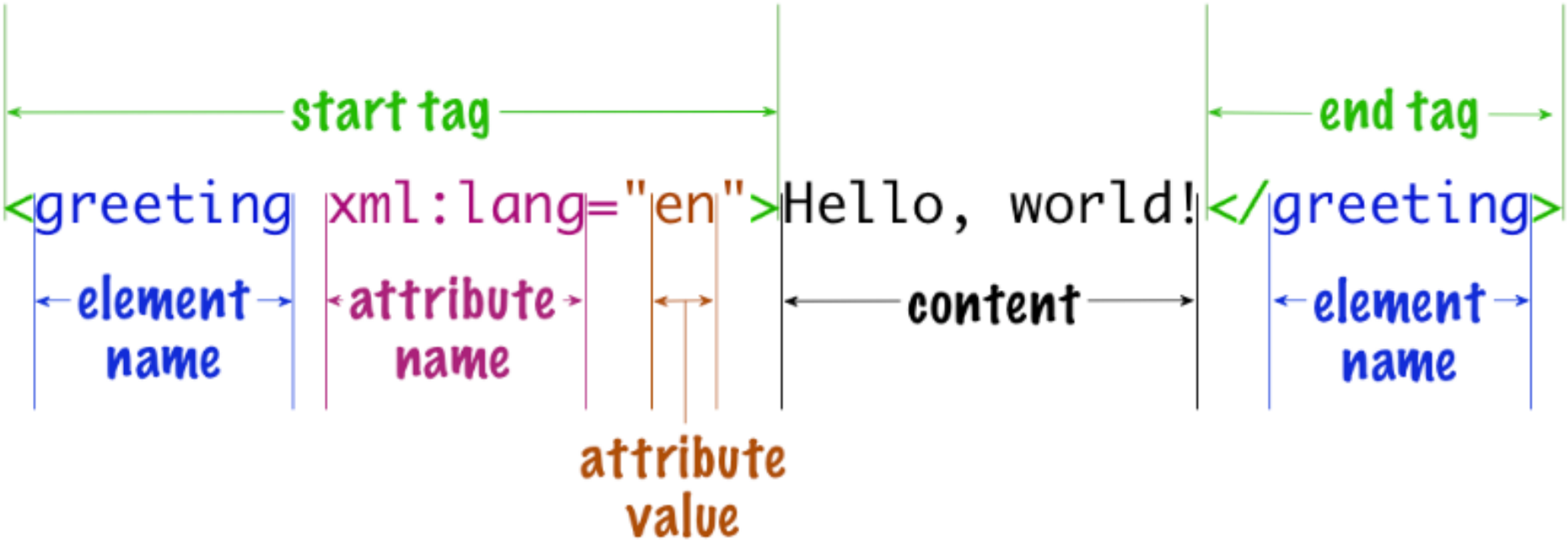
XML Elements & Tags

- 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>**



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>
```


XML Structure

Workshop:

Instructors:

KAK

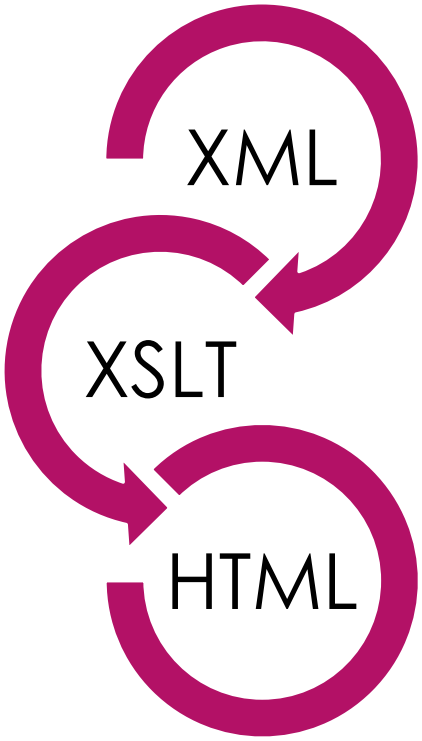
Participants:

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>
```



The Input Document

```
<?xml version="1.0" encoding="UTF-8"?>
<paragraph>Hello, world!</paragraph>
```

The Output Document

```
<html>
    <head>
        <title>A Short Test Document</title>
    </head>
    <body>
        <div>
            <p>
                ????
            </p>
        </div>
    </body>
</html>
```

The Stylesheet

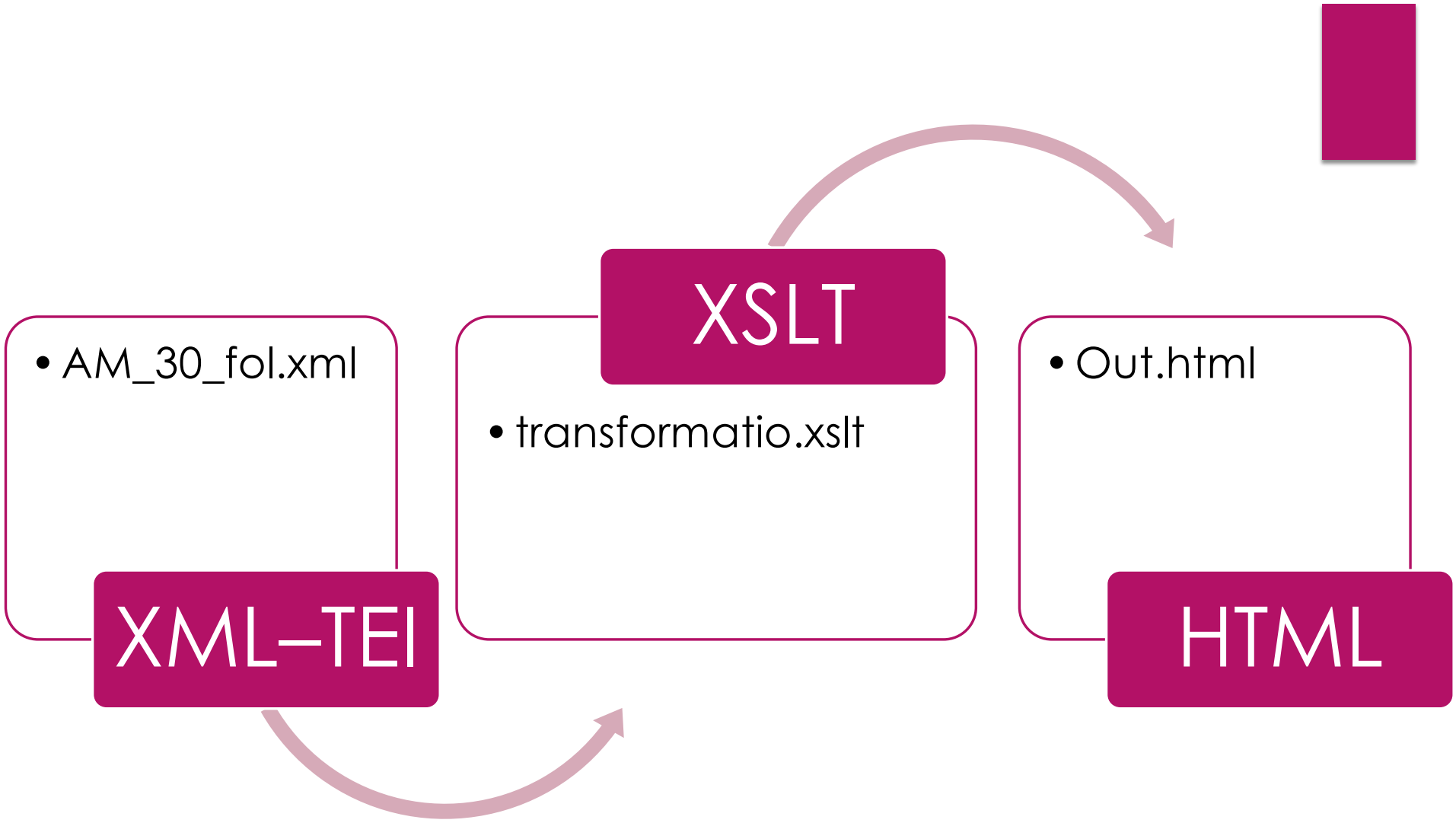
```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="1.0">

  <xsl:template match="paragraph">
    <html>
      <head>
        <title>A Short Test Document</title>
      </head>
      <body>
        <div>
          <p>
            <xsl:apply-templates/>
          </p>
        </div>
      </body>
    </html>
  </xsl:template>

</xsl:stylesheet>
```

Exercise Week1_Ex3 (https://tinyurl.com/Wk1Ex)

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


• AM_30_fol.xml
XML–TEI
XSLT
• transformatio.xslt
• Out.html
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*

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 l
23                   gathered into five quires of six conjoint l
24                   the last quire, with consists of four conj
25                   made of one type of relatively thick laid p
26                   AM 30 fol. contains two main texts dealing
27                   Slavs.</summary>
28                <msItem>
29                   <title>Chronica Slavorum</title>
30                   <author>Helmold of Bosau</author>
31                </msItem>
32                <msItem>
33                   <title>Chronica Slavorum</title>
```

Text Grid Author

Our XML file:
AM_30_fol.xml

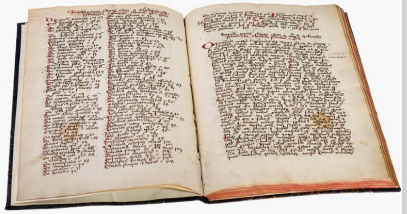
COPENHAGEN, DEN ARNAMAGNÆANSKE SAMLING, AM 30 FOL.

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*



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