TEI-XML Basics

XML vs TEI

XML (Extensible Markup Language) is a coding language that allows us to store, transmit, and reconstruct data in a way that is both human-readable and machine-readable. For our purposes, this means we can markup our data (i.e., the text) with notations. TEI (Text Encoding Initiative) is a flavor of XML, which means it runs on the same structure as XML, but it includes elements that are specific for the needs of marking up texts.

Elements (also known as tags) in XML and TEI need to be formed correctly. If they don't, then the code will not be valid and will generate error codes.

- ❖ Generally speaking, elements require two parts: an opening element and a closing element. A closing element will come second and will have a backslash at the beginning of the element.
    - ➢ Ex. <persName>Henry</persName>.
- ❖ If an element does not need to be wrapping a text, sometimes it will only have one element. When this is the case, the backslash will be included at the end of the element.
    - ➢ Ex. <space rend="indent1"/>
- ❖ The majority of elements that we will find in our TEI-XML documents are transformed from the paragraph styles and character styles we use in InDesign. When we are auditing our TEI-XML documents, it helps to have the corresponding InDesign document open.
    - ➢ It is important to note that, when elements are being drawn from **paragraph styles**, a new element will open after every end of paraph mark. This is why there are three separate opening and closing elements for <p style="body"> when we have three paragraphs in a row all styled with the Body paragraph style.
- ❖ Tags often need to be in a certain order in order for a document to be well-formed and validate.
    - ➢ Many tags are part of a parent-child relationship.
        - ■ Parent tags are containers that can hold a variety of smaller tags.
        - ■ Child tags are smaller tags that need to be surrounded by larger parent tags.
        - ■ Most of these parent-child relationships are established in our code by our transformation script.
    - ➢ We also nest tags that don't necessarily have parent-child relationships within each other.
        - ■ Move from general to specific: an element that applies to a full line of text should be on the outside, and an element that applies to only one word in that line should be within.
            - ● Example: Text that should be bold but then there is a Person!

- Render as <hi rend="bold">Text that should be bold but then there is a <persName>Person</persName>!</hi>

It is important to note that neither XML nor TEI are display languages. This means that any styling we need to do for the text to display properly on our website is connected to a CSS (cascading style sheet) where we define how we want styles to look.

Below, we'll go through some of the basics of what to expect from an XML document. For a (*very*) detailed overview of TEI, its structure, and all its elements, take a look through the TEI Guidelines available on tei-c.org.

Navigating an XML File

When you open an XML file in oXygen, the first few lines will typically say something like:

```
<?xml version="1.0" encoding="utf-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" version="3.3.0">
```

These first two lines define the entire document's structure. They indicate which version of XML we are using, as well as which version of TEI we are using. Knowing which version we are following allows the program to check whether the rest of the content validates.

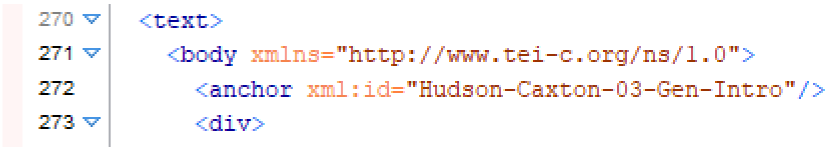
The rest of the document is structured in two overarching containers: the <teiHeader> and <text> elements.

<teiHeader>

The TEI header holds all important metadata specific to its particular XML file. This information is important for our metadata transparency. At this stage (intermediate, pre-launch), we do not need to audit this information.

<text>

The text container holds all major information and data for the XML document. The opening <text> element will always be followed by <body xmlns="http://www.tei-c.org/ns/1.0">, a file-specific <anchor> element, and a <div> element.



```
<text>
  <body xmlns="http://www.tei-c.org/ns/1.0">
    <anchor xml:id="Hudson-Caxton-03-Gen-Intro"/>
    <div>
```

The <div> elements are text division elements. They indicate large subdivisions within the <body> of the <text> elements. We use them primarily to separate the contents of the text.

Usually, there will be two main <div> elements: one for the main content and one for footnotes. Exceptions to this are prose texts with chapters, where each chapter receives its own <div> along with its own <div> for footnotes, and facing page texts. Each <div> element should receive a type attribute with a one word description of the type of content to be found within.

- Ex. <div type="footnotes">

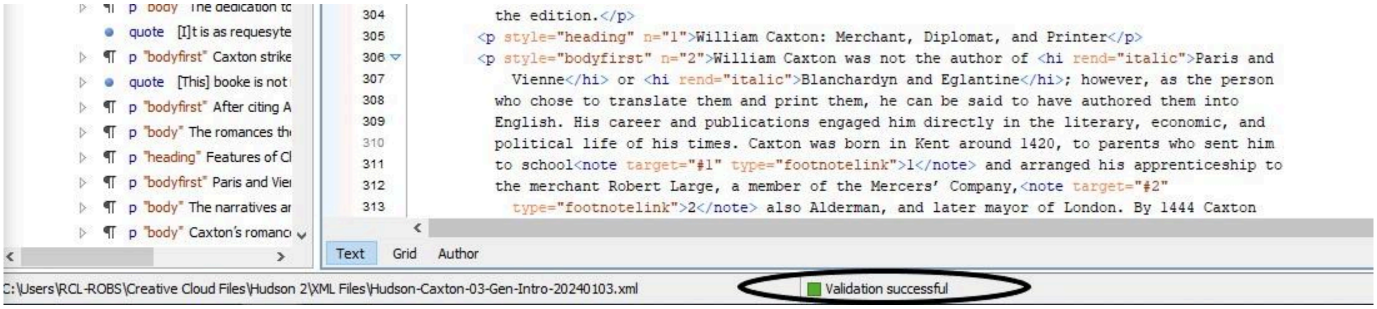
Using oXygen XML Editor

We use the program oXygen XML Editor to view, audit, and edit our TEI-XML files. This program is the standard in the field for editing XML files. It has a lot of functionality to it, which comes with many, many buttons. We only need to use a few of these buttons for our purposes.

Validation

If an XML file is correct and correctly formed (aka "well-formed"), then it validates. You can check this a couple ways in oXygen.

- oXygen displays a status update on whether a document validates as you make changes within it. Look at the bar along the bottom of the screen.
- The status update on validation lives in the middle of the bar. There will be one of two results.
    - A green square and it will say Validation Successful.

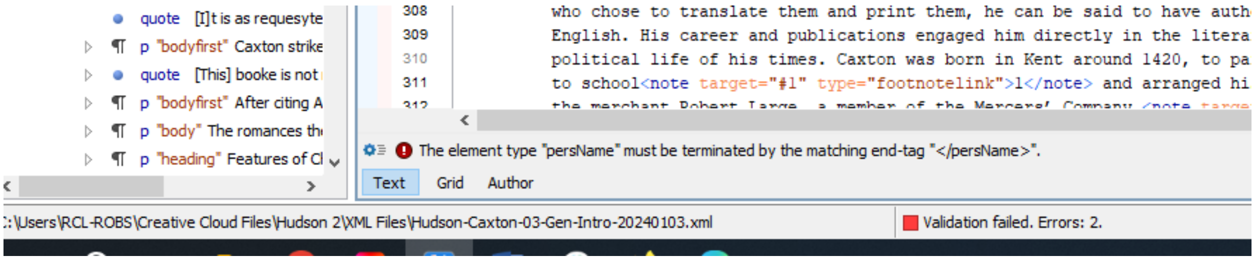

p "body" The dedication to
quote [I]t is as requesyte
p "bodyfirst" Caxton strike
quote [This] booke is not
p "bodyfirst" After citing A
p "body" The romances th
p "heading" Features of Cl
p "bodyfirst" Paris and Vie
p "body" The narratives ar
p "body" Caxton's romanc
Text
Grid
Author

```
   the edition.</p>
 <p style="heading" n="1">William Caxton: Merchant, Diplomat, and Printer</p>
 <p style="bodyfirst" n="2">William Caxton was not the author of <hi rend="italic">Paris and
     Vienne</hi> or <hi rend="italic">Blanchardyn and Eglantine</hi>; however, as the person
   who chose to translate them and print them, he can be said to have authored them into
   English. His career and publications engaged him directly in the literary, economic, and
   political life of his times. Caxton was born in Kent around 1420, to parents who sent him
   to school<note target="#1" type="footnotelink">1</note> and arranged his apprenticeship to
   the merchant Robert Large, a member of the Mercers' Company,<note target="#2"
     type="footnotelink">2</note> also Alderman, and later mayor of London. By 1444 Caxton
```

C:\Users\RCL-ROBS\Creative Cloud Files\Hudson 2\XML Files\Hudson-Caxton-03-Gen-Intro-20240103.xml

🟩 Validation successful

-
    - A red square that says Validation failed, followed by the number of errors in the document. There will also throw up an alert along the bottom of the screen detailing what the error is.


quote [I]t is as requesyte
p "bodyfirst" Caxton strike
quote [This] booke is not
p "bodyfirst" After citing A
p "body" The romances th
p "heading" Features of Cl
Text
Grid
Author

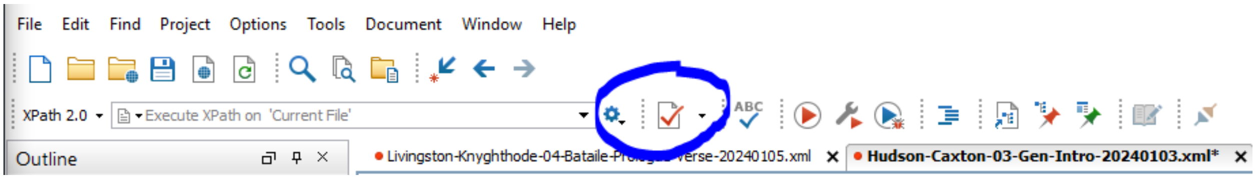
```
who chose to translate them and print them, he can be said to have auth
English. His career and publications engaged him directly in the litera
political life of his times. Caxton was born in Kent around 1420, to pa
to school<note target="#1" type="footnotelink">1</note> and arranged hi
```

The element type "persName" must be terminated by the matching end-tag "</persName>".

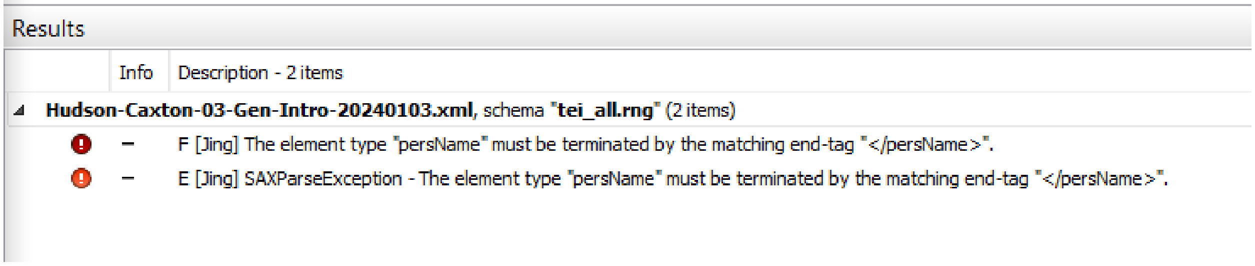
:\Users\RCL-ROBS\Creative Cloud Files\Hudson 2\XML Files\Hudson-Caxton-03-Gen-Intro-20240103.xml

🟥 Validation failed. Errors: 2.

- ❖ You can manually check the validation of the file by clicking the button that looks like a document with a red check mark:


File
Edit
Find
Project
Options
Tools
Document
Window
Help
XPath 2.0
Execute XPath on 'Current File'
Outline
Livingston-Knyghthode-04-Bataile-Prologue-Verse-20240105.xml
Hudson-Caxton-03-Gen-Intro-20240103.xml*

- ❖ This will pull up a Results bar along the bottom of the window. This will detail what errors there are, as well as where they exist.

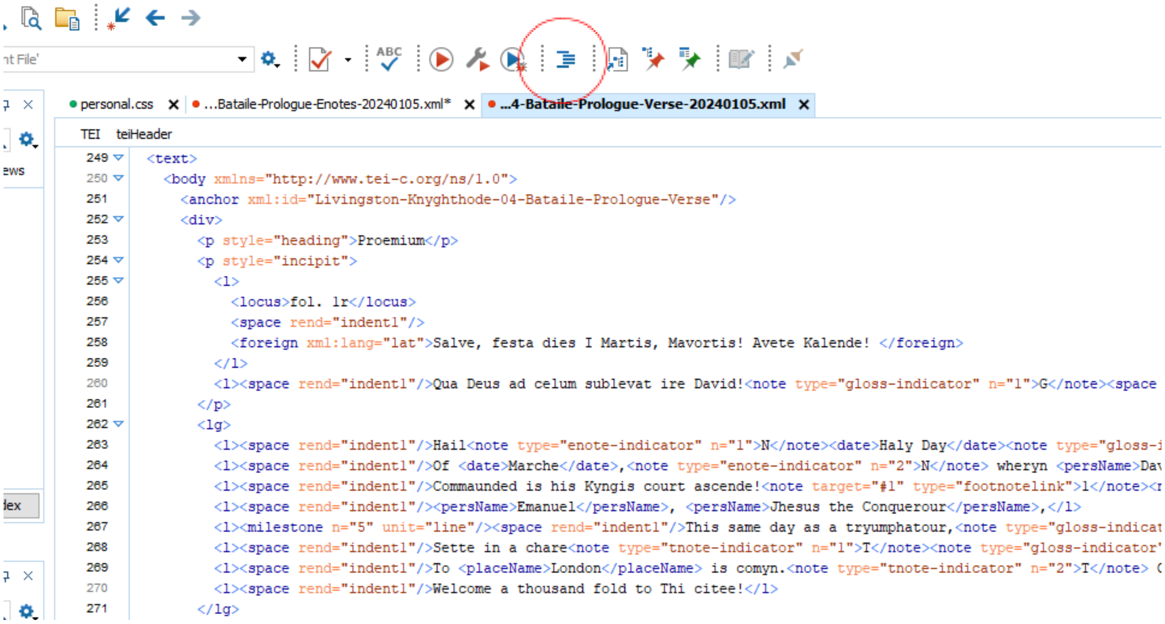


Hudson-Caxton-03-Gen-Intro-20240103.xml, schema "**tei_all.rng**" (2 items)

```
–    F [Jing] The element type "persName" must be terminated by the matching end-tag "</persName>".
–    E [Jing] SAXParseException - The element type "persName" must be terminated by the matching end-tag "</persName>".
```

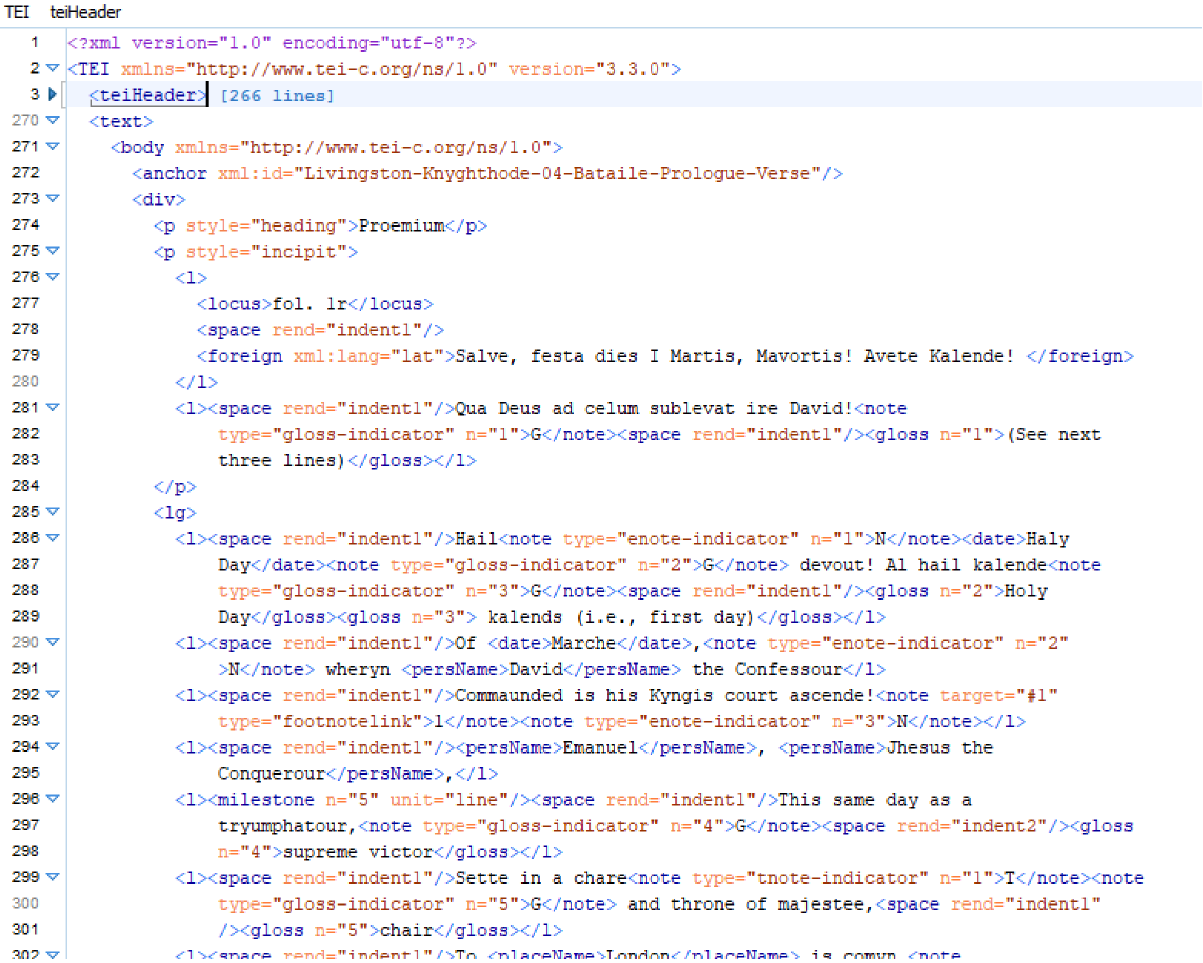
Format and Indent

When you first open an XML document, it prioritizes putting each element on its own line. This makes it good for seeing the structure quickly, but it's very difficult for us to navigate and audit. Instead, we'll want to change the way the document looks in oXygen. Click the Format and Indent button – it looks like a list with indents – in the top menu.


nt File'
personal.css
...Bataile-Prologue-Enotes-20240105.xml*
...4-Bataile-Prologue-Verse-20240105.xml
TEI
teiHeader

```
<text>
  <body xmlns="http://www.tei-c.org/ns/1.0">
    <anchor xml:id="Livingston-Knyghthode-04-Bataile-Prologue-Verse"/>
    <div>
      <p style="heading">Proemium</p>
      <p style="incipit">
        <l>
          <locus>fol. 1r</locus>
          <space rend="indent1"/>
          <foreign xml:lang="lat">Salve, festa dies I Martis, Mavortis! Avete Kalende! </foreign>
        </l>
        <l><space rend="indent1"/>Qua Deus ad celum sublevat ire David!<note type="gloss-indicator" n="1">G</note><space
      </p>
      <lg>
        <l><space rend="indent1"/>Hail<note type="enote-indicator" n="1">N</note><date>Haly Day</date><note type="gloss-i
        <l><space rend="indent1"/>Of <date>Marche</date>,<note type="enote-indicator" n="2">N</note> wheryn <persName>Da
        <l><space rend="indent1"/>Commaunded is his Kyngis court ascende!<note target="#1" type="footnotelink">1</note><
        <l><space rend="indent1"/><persName>Emanuel</persName>, <persName>Jhesus the Conquerour</persName>,</l>
        <l><milestone n="5" unit="line"/><space rend="indent1"/>This same day as a tryumphatour,<note type="gloss-indicat
        <l><space rend="indent1"/>Sette in a chare<note type="tnote-indicator" n="1">T</note><note type="gloss-indicator"
        <l><space rend="indent1"/>To <placeName>London</placeName> is comyn.<note type="tnote-indicator" n="2">T</note> (
        <l><space rend="indent1"/>Welcome a thousand fold to Thi citee!</l>
      </lg>
```

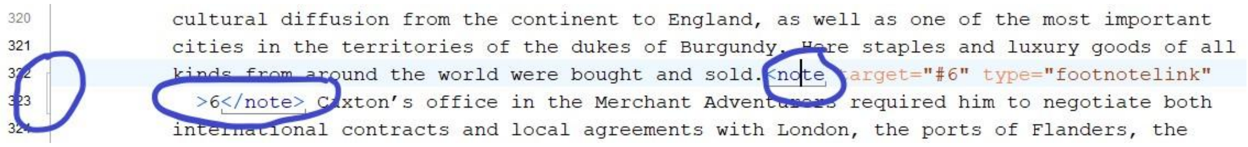
Then, the document will look like this:



```
<?xml version="1.0" encoding="utf-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" version="3.3.0">
  <teiHeader> [266 lines]
  <text>
    <body xmlns="http://www.tei-c.org/ns/1.0">
      <anchor xml:id="Livingston-Knyghthode-04-Bataile-Prologue-Verse"/>
      <div>
        <p style="heading">Proemium</p>
        <p style="incipit">
          <l>
            <locus>fol. 1r</locus>
            <space rend="indent1"/>
            <foreign xml:lang="lat">Salve, festa dies I Martis, Mavortis! Avete Kalende! </foreign>
          </l>
          <l><space rend="indent1"/>Qua Deus ad celum sublevat ire David!<note
              type="gloss-indicator" n="1">G</note><space rend="indent1"/><gloss n="1">(See next
              three lines)</gloss></l>
        </p>
        <lg>
          <l><space rend="indent1"/>Hail<note type="enote-indicator" n="1">N</note><date>Haly
              Day</date><note type="gloss-indicator" n="2">G</note> devout! Al hail kalende<note
              type="gloss-indicator" n="3">G</note><space rend="indent1"/><gloss n="2">Holy
              Day</gloss><gloss n="3"> kalends (i.e., first day)</gloss></l>
          <l><space rend="indent1"/>Of <date>Marche</date>,<note type="enote-indicator" n="2"
              >N</note> wheryn <persName>David</persName> the Confessour</l>
          <l><space rend="indent1"/>Commaunded is his Kyngis court ascende!<note target="#1"
              type="footnotelink">1</note><note type="enote-indicator" n="3">N</note></l>
          <l><space rend="indent1"/><persName>Emanuel</persName>, <persName>Jhesus the
              Conquerour</persName>,</l>
          <l><milestone n="5" unit="line"/><space rend="indent1"/>This same day as a
              tryumphatour,<note type="gloss-indicator" n="4">G</note><space rend="indent2"/><gloss
              n="4">supreme victor</gloss></l>
          <l><space rend="indent1"/>Sette in a chare<note type="tnote-indicator" n="1">T</note><note
              type="gloss-indicator" n="5">G</note> and throne of majestee,<space rend="indent1"
              /><gloss n="5">chair</gloss></l>
          <l><space rend="indent1"/>To <placeName>London</placeName> is comyn <note
```

Clicking on tags to see starting and ending elements

A quick way to see where elements open and close is by clicking in one of the tags. When you click into a tag, the entirety of the element will be underlined and a bracket will appear in the left margin, connecting the opening and closing elements.



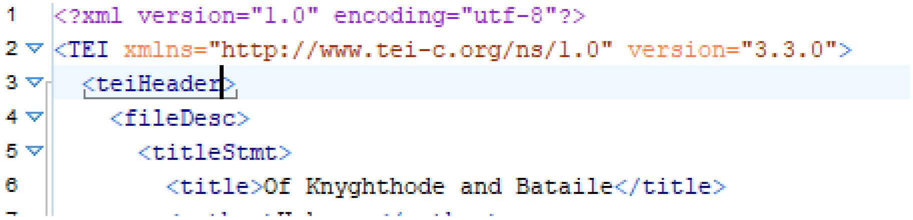
```
          cultural diffusion from the continent to England, as well as one of the most important
          cities in the territories of the dukes of Burgundy. Here staples and luxury goods of all
          kinds from around the world were bought and sold.<note target="#6" type="footnotelink"
            >6</note> Caxton’s office in the Merchant Adventurers required him to negotiate both
          international contracts and local agreements with London, the ports of Flanders, the
```

Collapsing elements

Once XML files are formatted and indented, they can be very lengthy. It can be easier to navigate the documents by collapsing down larger elements.

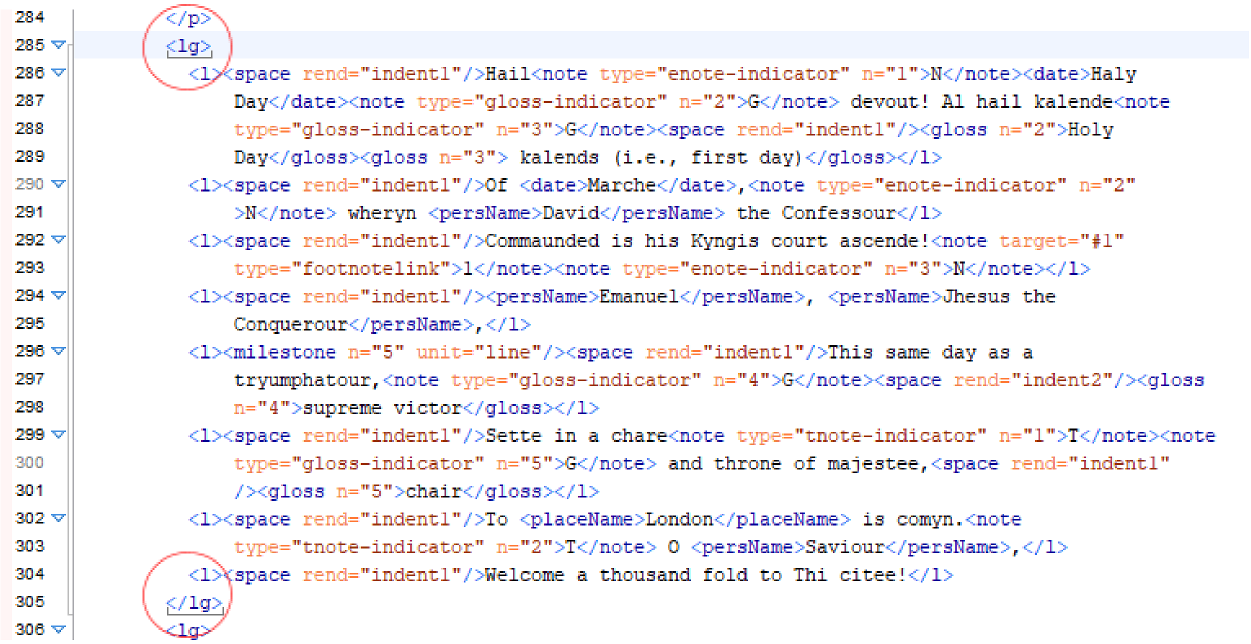
- ❖ Along the left margin is a line count for the code. Note the arrows beside some of the numbers. These indicate tags that have closing elements on a different line.

- ❖ Click the down arrow to collapse an entire section so it's not in your way.
  - ➢ We recommend you do this for the <teiHeader>, since it won't need to be audited for this stage.
  - ➢ Click into <teiHeader> and notice how it is underlined and a bracket appears, stretching far down the page. You can use the **down arrow** next to line 3 to **collapse** this entire header section so it's not in your way.



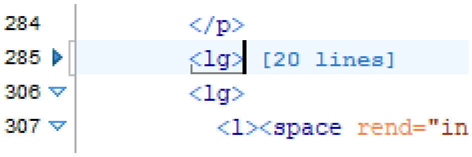
```
1   <?xml version="1.0" encoding="utf-8"?>
2 ▽ <TEI xmlns="http://www.tei-c.org/ns/1.0" version="3.3.0">
3 ▽   <teiHeader>
4 ▽     <fileDesc>
5 ▽       <titleStmt>
6           <title>Of Knyghthode and Bataile</title>
```

- Ex: The cursor is in the <lg> on XML line 285, so it appears with a small black bracket beneath it. This line group closes on XML line 305, so the </lg> is also bracketed there. To the left, these lines are highlighted by another small black bracket stretching from line 285 to 305.



```
284         </p>
285 ▽       <lg>
286 ▽         <l><space rend="indent1"/>Hail<note type="enote-indicator" n="1">N</note><date>Haly
287               Day</date><note type="gloss-indicator" n="2">G</note> devout! Al hail kalende<note
288               type="gloss-indicator" n="3">G</note><space rend="indent1"/><gloss n="2">Holy
289               Day</gloss><gloss n="3"> kalends (i.e., first day)</gloss></l>
290 ▽         <l><space rend="indent1"/>Of <date>Marche</date>,<note type="enote-indicator" n="2"
291               >N</note> wheryn <persName>David</persName> the Confessour</l>
292 ▽         <l><space rend="indent1"/>Commaunded is his Kyngis court ascende!<note target="#1"
293               type="footnotelink">1</note><note type="enote-indicator" n="3">N</note></l>
294 ▽         <l><space rend="indent1"/><persName>Emanuel</persName>, <persName>Jhesus the
295               Conquerour</persName>,</l>
296 ▽         <l><milestone n="5" unit="line"/><space rend="indent1"/>This same day as a
297               tryumphatour,<note type="gloss-indicator" n="4">G</note><space rend="indent2"/><gloss
298               n="4">supreme victor</gloss></l>
299 ▽         <l><space rend="indent1"/>Sette in a chare<note type="tnote-indicator" n="1">T</note><note
300               type="gloss-indicator" n="5">G</note> and throne of majestee,<space rend="indent1"
301               /><gloss n="5">chair</gloss></l>
302 ▽         <l><space rend="indent1"/>To <placeName>London</placeName> is comyn.<note
303               type="tnote-indicator" n="2">T</note> O <persName>Saviour</persName>,</l>
304           <l><space rend="indent1"/>Welcome a thousand fold to Thi citee!</l>
305       </lg>
306 ▽     <lg>
```

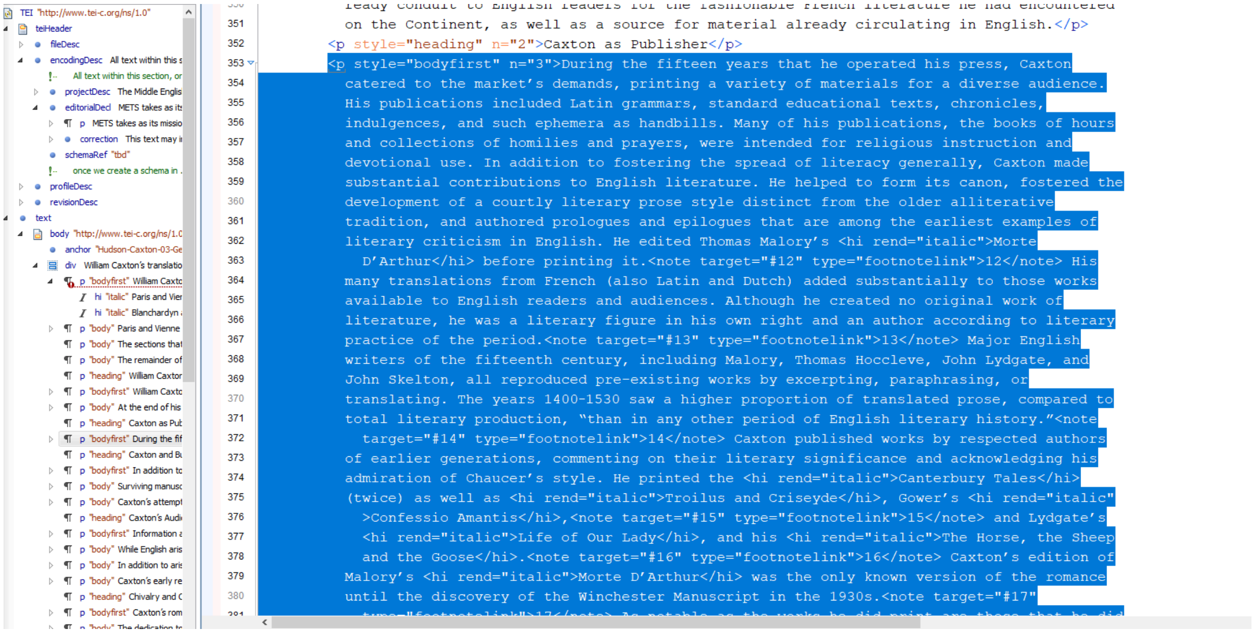
- When you collapse this line group, 285 will collapse and the next visible line will be 306. oXygen also helpfully tells you how many lines you've collapsed, so look for the bracketed [x number of lines] in case you fear things have vanished.



```
284         </p>
285 ▶       <lg> [20 lines]
306 ▽       <lg>
307 ▽         <l><space rend="in
```

N.B.: If you are having trouble navigating through the main elements in an XML file, it helps to have the Outline window open.

- If it's not already open, then click Window > Show View > Outline
- An outline will appear along the left side of your screen.
- When you click on an element in the outline view, then the main display of the XML file will jump to the element you have selected, as well as highlight the entirety of the element in blue.


TEI "http://www.tei-c.org/ns/1.0"
teiHeader
fileDesc
encodingDesc All text within this s
!-- All text within this section, or
projectDesc The Middle Englis
editorialDecl METS takes as its
¶ p METS takes as its missio
correction This text may i
schemaRef "tbd"
!-- once we create a schema in .
profileDesc
revisionDesc
text
body "http://www.tei-c.org/ns/1.0
anchor "Hudson-Caxton-03-Ge
div William Caxton's translatio
¶ p "bodyfirst" William Caxto
hi "italic" Paris and Vien
hi "italic" Blanchardyn a
¶ p "body" Paris and Vienne
¶ p "body" The sections that
¶ p "body" The remainder of
¶ p "heading" William Caxton
¶ p "bodyfirst" William Caxto
¶ p "body" At the end of his
¶ p "heading" Caxton as Pub
¶ p "bodyfirst" During the fif
¶ p "heading" Caxton and Bu
¶ p "bodyfirst" In addition to
¶ p "body" Surviving manusc
¶ p "body" Caxton's attempt
¶ p "heading" Caxton's Audi
¶ p "bodyfirst" Information a
¶ p "body" While English aris
¶ p "body" In addition to aris
¶ p "body" Caxton's early re
¶ p "heading" Chivalry and C
¶ p "bodyfirst" Caxton's rom
¶ p "body" The dedication to

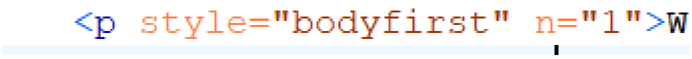
```
      ready conduit to English readers for the fashionable French literature he had encountered
      on the Continent, as well as a source for material already circulating in English.</p>
    <p style="heading" n="2">Caxton as Publisher</p>
    <p style="bodyfirst" n="3">During the fifteen years that he operated his press, Caxton
      catered to the market's demands, printing a variety of materials for a diverse audience.
      His publications included Latin grammars, standard educational texts, chronicles,
      indulgences, and such ephemera as handbills. Many of his publications, the books of hours
      and collections of homilies and prayers, were intended for religious instruction and
      devotional use. In addition to fostering the spread of literacy generally, Caxton made
      substantial contributions to English literature. He helped to form its canon, fostered the
      development of a courtly literary prose style distinct from the older alliterative
      tradition, and authored prologues and epilogues that are among the earliest examples of
      literary criticism in English. He edited Thomas Malory's <hi rend="italic">Morte
        D'Arthur</hi> before printing it.<note target="#12" type="footnotelink">12</note> His
      many translations from French (also Latin and Dutch) added substantially to those works
      available to English readers and audiences. Although he created no original work of
      literature, he was a literary figure in his own right and an author according to literary
      practice of the period.<note target="#13" type="footnotelink">13</note> Major English
      writers of the fifteenth century, including Malory, Thomas Hoccleve, John Lydgate, and
      John Skelton, all reproduced pre-existing works by excerpting, paraphrasing, or
      translating. The years 1400-1530 saw a higher proportion of translated prose, compared to
      total literary production, "than in any other period of English literary history."<note
        target="#14" type="footnotelink">14</note> Caxton published works by respected authors
      of earlier generations, commenting on their literary significance and acknowledging his
      admiration of Chaucer's style. He printed the <hi rend="italic">Canterbury Tales</hi>
      (twice) as well as <hi rend="italic">Troilus and Criseyde</hi>, Gower's <hi rend="italic"
        >Confessio Amantis</hi>,<note target="#15" type="footnotelink">15</note> and Lydgate's
        <hi rend="italic">Life of Our Lady</hi>, and his <hi rend="italic">The Horse, the Sheep
        and the Goose</hi>.<note target="#16" type="footnotelink">16</note> Caxton's edition of
      Malory's <hi rend="italic">Morte D'Arthur</hi> was the only known version of the romance
      until the discovery of the Winchester Manuscript in the 1930s.<note target="#17"
        type="footnotelink">17</note> As notable as the works he did print are those that he did
```

<ins>The color coding system</ins>

oXygen color-codes information to make it easier to parse the code at a glance. You will typically see the following colors in your XML documents:

- Black: any content/data contained by the code
- Dark blue: main elements used in TEI-XML
- Orange: attributes that help define the elements
- Dark red: definitions for the attributes
- Green: a comment made within the code. This is essentially a note-to-self that doesn't display.
- Light pea-green: an escaped character.

Ex.



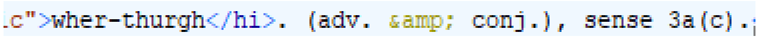
```
<p style="bodyfirst" n="1">W
```

In this case, the “p” is the element, the orange “style=” and “n=” are attributes, and “bodyfirst” and “1” are the further definitions for the attributes.

Escaped characters

Certain characters, such as ampersands, are command characters for oXygen, which are used to signal the code to perform some sort of action. In order for these characters to display in the content, we have to use the unicode character code. Some diacritics also display through escaped characters. Anything represented by unicodes will display as a light pea-green color.

Ex:



```
.c">wher-thurgh</hi>. (adv. &amp; conj.), sense 3a(c).
```