GENEALOGIES OF KNOWLEDGE

How to clean up documents and prepare them for the pre-modern corpus

Purpose: To produce XML text and header files that correspond to the contents of the book to be incorporated into the premodern corpus, and that validate against the DTD header and text files.

BUILDING THE PREMODERN CORPUS

In most cases, you will be working with files that have already been marked up in some flavour of XML. The main tasks for the premodern corpus, then, are to (1) create a header file that validates against the GoK header DTD, and (2) to modify a pre-existing XML file (usually in TEI format) so that it validates against the GoK text DTD.

CREATING THE GoK HEADER FILE

Header files for premodern texts are two types (a) source language texts (usually, but not always, Greek); and (b) translated texts. The structure of the GoK header will differ depending on which type of text you are annotating.

PREPARING THE DOCUMENT WITH OXYGEN XML EDITOR (v. 16.0)

Using Oxygen for annotation:

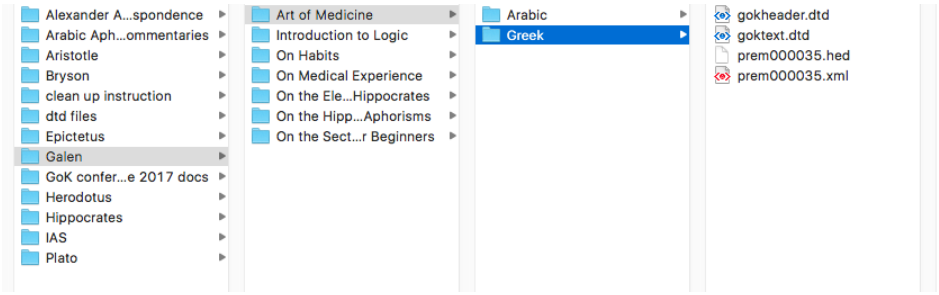
While Oxygen performs all the functions of jEdit, the former is able to display Right-to-Left languages such as Arabic.

If Oxygen XML Editor is not installed on your computer, the current version (v. 19 as 22 May 2017) can be downloaded here: https://www.oxygenxml.com/xml_editor/software_archive_editor.html

Linking the annotated texts to DTD:

Before beginning the annotation, the DTD header and text files must be placed in the same folder as the annotated text and header files. As long as the DTD files in the same folder as the text and header files, Oxygen will automatically VALIDATE the XML files (header and text) *as you work*.

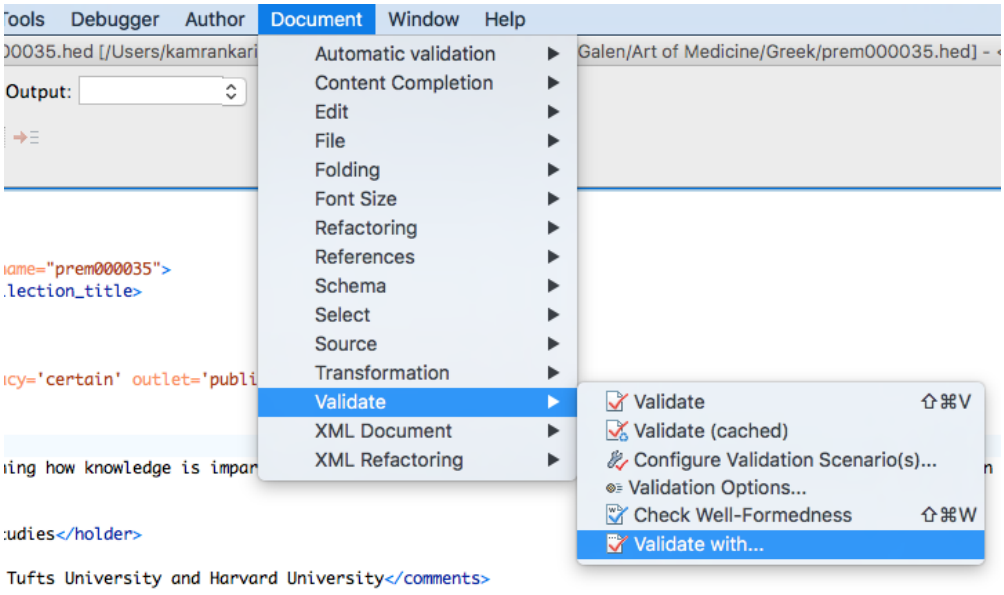
1. For example, if we assign Galen's *On the Art of Medicine* the text id prem000035, the file structure might look like the following:



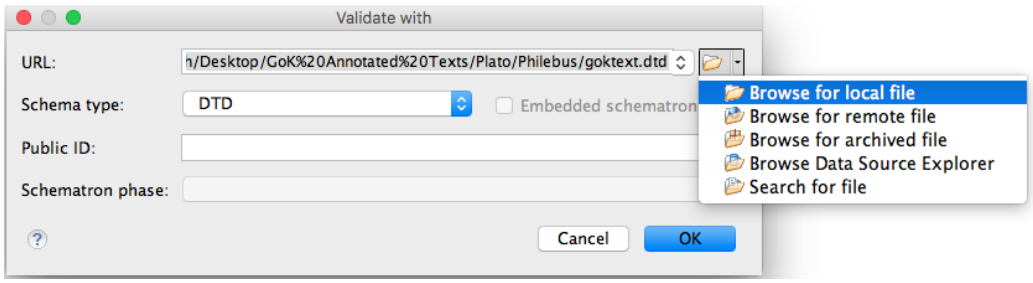
Thus, the header file must be named prem000035.hed and the annotated text must be named prem000035.xml.

2. Oxygen must then be directed to the DTD files against which we want to validate the annotated text (premn000035.xml) and the header file (prem000035.hed).

On the Oxygen menu bar: Select DOCUMENTS → VALIDATE → VALIDATE WITH …

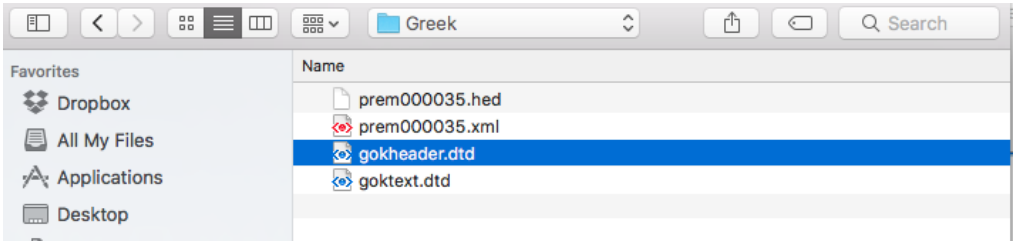


3. The next step is to associate the appropriate DTDs to the respective XML files. Thus, gokheader.dtd needs to be associated with prem000035.hed and goktext.dtd needs to be associated with prem000035.xml.

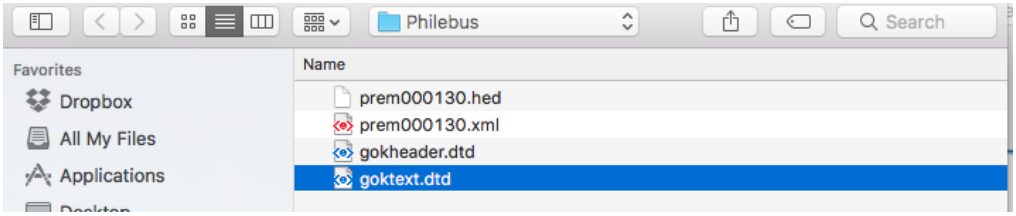


*From **Validate with menu***: Click on the **Folder icon** and Select **BROWSE FOR LOCAL FILE**.

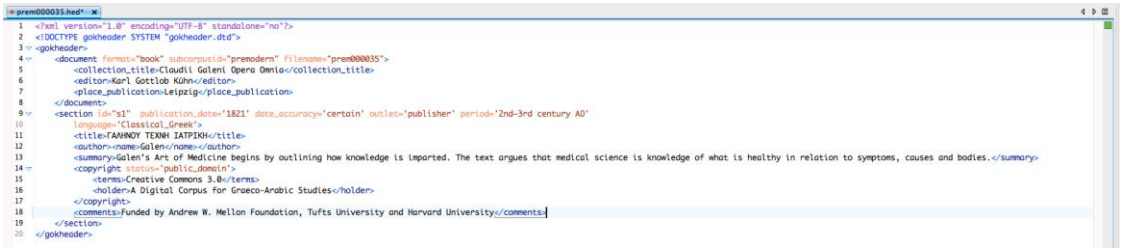
To VALIDATE prem000035.hed, associate gokheader.dtd with it.



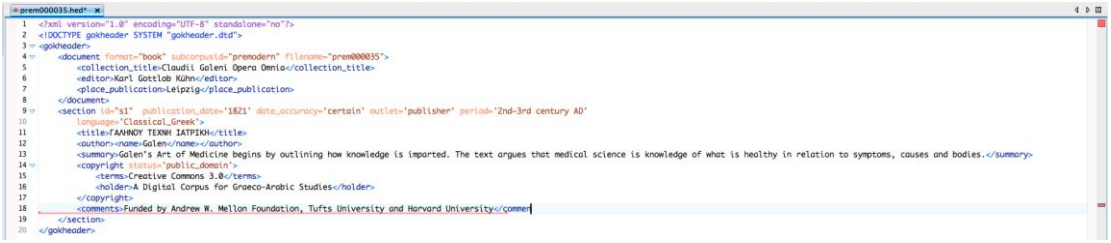
To VALIDATE prem000035.xml, associate goktext.dtd with it.



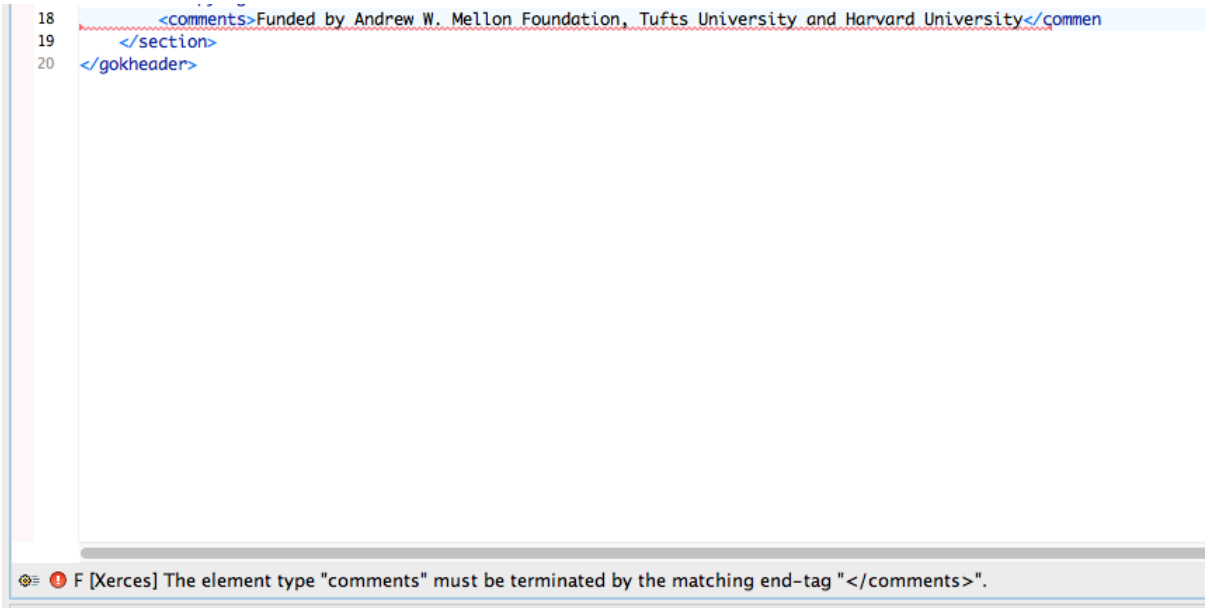
When the file is VALID according to the DTD associated with it, Oxygen indicates this by showing a green box in the right sign of the window.



When the file is not VALID according the DTD associated with it, the box turns red:



Oxygen also provides the reason for the error at the bottom of the editor window:



Finally, it is often helpful to have a printed copy of the 'goktext.dtd' and 'gokheader.dtd' files to hand when adding the metadata. These documents include all the possible metadata tags that you can use to annotate the text and it is against these .dtds that Oxygen will check for errors.

a. Source Language Texts (Greek)

Header files tell the reader what the text is by presenting a list of features that uniquely characterise the text. These features are called ELEMENTS. In the header file below, Galen's *On the Art of Medicine* has ELEMENTS such as <title> and <place_publication>, the values of which tell the reader what the text is. Header files for source texts will typically have something like the following structure. *In most cases, it will be possible to cut and paste the lines below into the header file, simply modifying the values assigned ELEMENTS and ATTRIBUTES for the current document.*

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
  <document format="book" subcorpusid="premodern" filename="prem000035">
    <collection_title>Claudii Galeni Opera Omnia</collection_title>
    <editor>Karl Gottlob Kühn</editor>
    <place_publication>Leipzig</place_publication>
     </document><section id="s1"  publication_date='1821' date_accuracy='certain'
   outlet='publisher' period='2nd–3rd century AD'
        language='Classical_Greek'>
        <title>ΓΑΛΗΝΟΥ ΤΕΧΝΗ ΙΑΤΡΙΚΗ</title>

   <summary>Galen's Art of Medicine begins by outlining how knowledge is imparted.
   The text argues that medical science is knowledge of what is healthy in relation to
```

```
symptoms, causes and bodies.</summary>
      <author><name>Galen</name></author>
      <copyright status='public_domain'>
        <terms>Creative Commons 3.0</terms>
        <holder>A Digital Corpus for Graeco-Arabic Studies</holder>
      </copyright>
      <comments>Funded by Andrew W. Mellon Foundation, Tufts University and Harvard
University</comments>
    </section>
</gokheader>
```

The first two lines of the header file

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
```

must be the same in all header files for Genealogies of Knowledge.

All the ELEMENTS in this header file must be opened and closed. You will notice, for example, that <gokheader> opens the ELEMENT and </gokheader> closes it at the end of the header file.

ELEMENTS can have ATTRIBUTES. For example, the ELEMENT <document> has the ATTRIBUTE @book, and these attributes are assigned values using an assignment operator = and the value assigned is placed between quotation marks. Thus, within the ELEMENT <section> the ATTRIBUTE @publication_date is assigned the value 1821 by writing: publication_date="1821".

@filename is a unique id, which for the premodern corpus has the form premXXXXXX, where the Xs are digits. For example, in the header file the value prem000035 is assigned to Galen's text.

b. Translated Texts

The following header may serve as a template for the structure of most pre-modern translated texts, modifying the values of ATTRIBUTES and populating ELEMENTS according to the current document.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE gokheader SYSTEM "gokheader.dtd">
<gokheader>
  <document format="book" subcorpusid="premodern" filename="prem000039">
    <collection_title>Arabic Commentaries on the Hippocratic Aphorisms</collection_title>
    <editor>Taro Mimura</editor>
    <place_publication>Manchester</place_publication>
  </document>
  <section id="s1"  publication_date='2012–2017' date_accuracy='certain' outlet='website'
period='9th century AD'
    language='Classical_Greek'>
    <title>Tafsīr Ǧālīnūs li-Fuṣūl Abuqrāṭ</title>
```

<summary>Hippocrates' Aphorisms discuss the nature of medical experience, medical reasoning, prognosis, regimen and differences in nosology and aetiology based on gender, climate and geographical location. Galen's commentary expands on Hippocrates ideas and offers an underlying medical theory of humours, elements, system of prognosis, and rationales for Hippocrates' judgments. All this is supplemented by observations on human anatomy and an insistence on a scientific method for medicine that combines reasoning and observation.</summary>
<author><name>Galen</name></author>
<translation status="translation" language="Medieval_Arabic">
<translator certainty="certain">
<name>Ḥunayn ibn Isḥāq</name>
</translator>
<source date="2nd or 3rd century AD"><original_title>ΙΠΠΟΚΡΑΤΟΥΣ ΑΦΟΡΙΣΜΟΙ ΚΑΙ ΓΑΛΗΝΟΥ ΕΙΣ ΑΥΤΟΥΣ ΥΠΟΜΝΗΜΑΤΑ.</original_title></source>
</translation>
<copyright status='public_domain'>
<terms>Creative Commons BY</terms>
<holder>The University of Manchester</holder>
</copyright>
<comments>Edition was funded by the ERC. Galen's author TLG id is tlg0057. Aphorisms commentary is tlg0057.tlg.0092. This file is Book One of Galen's commentary.</comments>
</section>
</gokheader>

There are some similarities between the header for the source and translated text. Obviously, elements such as author will be the same; likewise, the summaries for the source and translated texts are normally identical.

There are, however, several differences between headers for source texts and translated texts. The <translation> ELEMENT will need to be populated with ATTRIBUTES such as @language, which may take several possible values (for example, "Medieval_Arabic," "Latin," "English"). In the case of pre-modern texts, the identity of the translator is often uncertain or even unknown. The ELEMENT translator must be assigned a value indicating how certain we are that the translator included in the <name> ELEMENT is, in fact, the translator.

One final remark. The ELEMENT <source> is an obligatory SUB-ELEMENT of <translation>. In addition to the SUB-ELEMENT <original_title>, it is also necessary to specify the date the original text was written. We specify this information by assigning the ATTRIBUTE @date a value. For pre-modern texts, often the value assigned to @date will the century in which the author was known to be alive.

CREATING THE GoK TEXT FILE

Owing to the fact that most of the pre-modern texts are typically marked-up in XML, you will need to remove or modify the mark-up that is not permitted by the GoK text DTD. The most efficient way of modifying pre-existing XML documents is to use "regular expressions" (hereafter: regex). See the introduction to using regex for the

Genealogies of Knowledge project: http://genealogiesofknowledge.net/genealogies-knowledge-corpus/#instructions

The following section will take you through the process of modifying a Greek text marked-up in TEI so that it validates against the GoK text DTD.

USING OXYGEN AND REGEX TO CLEAN UP TEXTS

Here is a segment of text marked-up in TEI that requires clean-up:

…

```
<body>
   <milestone unit="Book" n="1"/>
   <div n="1094a" type="bekker page">
      <milestone n="1" unit="bekker line"/>
      <head lang="en">βοοκ 1</head>
      <p>πᾶσα τέχνη καὶ πᾶσα μέθοδος, ὁμοίως δὲ πρᾶξίς τε καὶ προαίρεσις, ἀγαθοῦ τινὸς ἐφίεσθαι
δοκεῖ· διὸ καλῶς ἀπεφήναντο τἀγαθόν, οὗ πάντ’ ἐφίεται. διαφορὰ δέ τις φαίνεται τῶν τελῶν· τὰ μὲν γάρ
εἰσιν ἐνέργειαι, τὰ δὲ παρ’ αὐτὰς <milestone n="5" unit="bekker line"/>ἔργα τινά. ὧν δ’ εἰσὶ τέλη τινὰ
παρὰ τὰς πράξεις, ἐν τούτοις βελτίω πέφυκε τῶν ἐνεργειῶν τὰ ἔργα. πολλῶν δὲ πράξεων οὐσῶν καὶ
τεχνῶν καὶ ἐπιστημῶν πολλὰ γίνεται καὶ τὰ τέλη· ἰατρικῆς μὲν γὰρ ὑγίεια, ναυπηγικῆς δὲ πλοῖον,
στρατηγικῆς δὲ νίκη, οἰκονομικῆς δὲ πλοῦτος. ὅσαι <milestone n="10" unit="bekker line"/>δ’ εἰσὶ τῶν
τοιούτων ὑπὸ μίαν τινὰ δύναμιν, καθάπερ ὑπὸ τὴν ἱππικὴν χαλινοποιικὴ καὶ ὅσαι ἄλλαι τῶν ἱππικῶν
ὀργάνων εἰσίν, αὕτη δὲ καὶ πᾶσα πολεμικὴ πρᾶξις ὑπὸ τὴν στρατηγικήν, κατὰ τὸν αὐτὸν δὴ τρόπον
ἄλλαι ὑφ’ ἑτέρας· ἐν ἁπάσαις δὲ τὰ τῶν ἀρχιτεκτονικῶν τέλη πάντων <milestone n="15" unit="bekker
line"/>ἐστὶν αἱρετώτερα τῶν ὑπ’ αὐτά· τούτων γὰρ χάριν κἀκεῖνα διώκεται. διαφέρει δ’ οὐδὲν τὰς
ἐνεργείας αὐτὰς εἶναι τὰ τέλη τῶν πράξεων ἢ παρὰ ταύτας ἄλλο τι, καθάπερ ἐπὶ τῶν λεχθεισῶν
ἐπιστημῶν. εἰ δή τι τέλος ἐστὶ τῶν πρακτῶν ὃ δι’ αὐτὸ βουλόμεθα, τἆλλα δὲ διὰ τοῦτο, καὶ μὴ
<milestone n="20" unit="bekker line"/>πάντα δι’ ἕτερον αἱρούμεθα ‘πρόεισι γὰρ οὕτω γ’ εἰς ἄπειρον,
ὥστ’ εἶναι κενὴν καὶ ματαίαν τὴν ὄρεξιν’, δῆλον ὡς τοῦτ’ ἂν εἴη τἀγαθὸν καὶ τὸ ἄριστον. ἆρ’ οὖν καὶ πρὸς
τὸν βίον ἡ γνῶσις αὐτοῦ μεγάλην ἔχει ῥοπήν, καὶ καθάπερ τοξόται σκοπὸν ἔχοντες μᾶλλον ἂν
τυγχάνοιμεν τοῦ δέοντος; εἰ δ’ <milestone n="25" unit="bekker line"/>οὕτω, πειρατέον τύπῳ γε
περιλαβεῖν αὐτὸ τί ποτ’ ἐστὶ καὶ τίνος τῶν ἐπιστημῶν ἢ δυνάμεων. δόξειε δ’ ἂν τῆς κυριωτάτης καὶ
μάλιστα ἀρχιτεκτονικῆς. τοιαύτη δ’ ἡ πολιτικὴ φαίνεται· τίνας γὰρ εἶναι χρεὼν τῶν ἐπιστημῶν ἐν ταῖς
πόλεσι,
      </p>
```

…

1. Step one is to replace the TEI header with the DTD header. This involves replacing the TEI header with the following lines:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE goktext SYSTEM "goktext.dtd">
```

2. This also involves ensuring that the text is inserted into the <goktext> and <section> ELEMENTS. Note that in most cases, @id will be assigned the value s1.

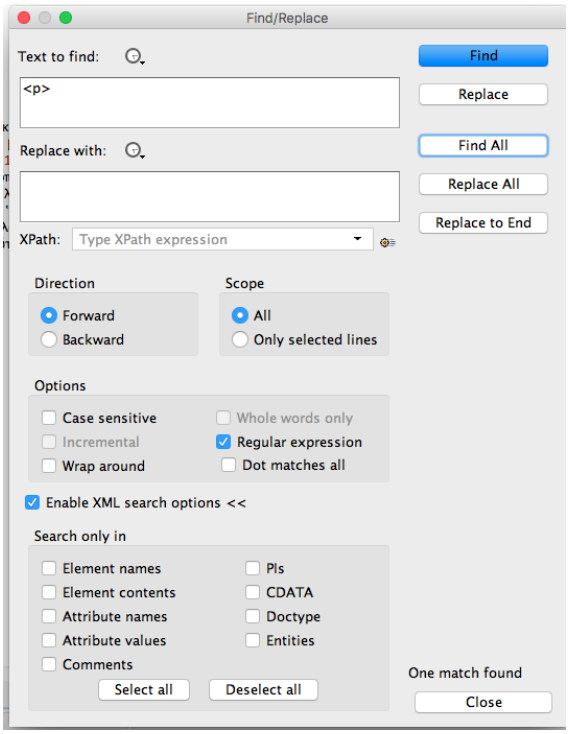
```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE goktext SYSTEM "goktext.dtd">
```

```
<goktext>
  <section id="s1">

<insert your text here>

  </section>
</goktext>
```

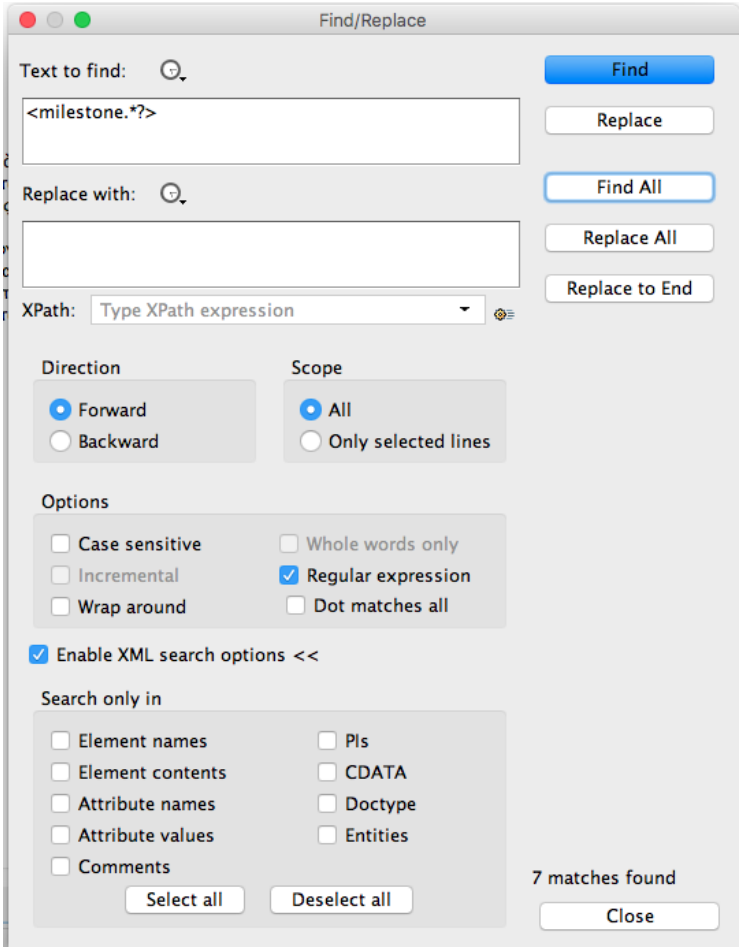
3. The next step is to ensure that all the ELEMENTS that are not allowed by the current DTD either are **removed** or **replaced** by ELEMENTS that are allowed whilst preserving the information contained in them.

    a. Removed—In the above sample, the ELEMENTS <p>, <body>, <milestone> and <head> must be removed because they are not allowed by the text DTD and they do not contain any useful information. For example,

        i. To remove <p>, use the following find an replace command:

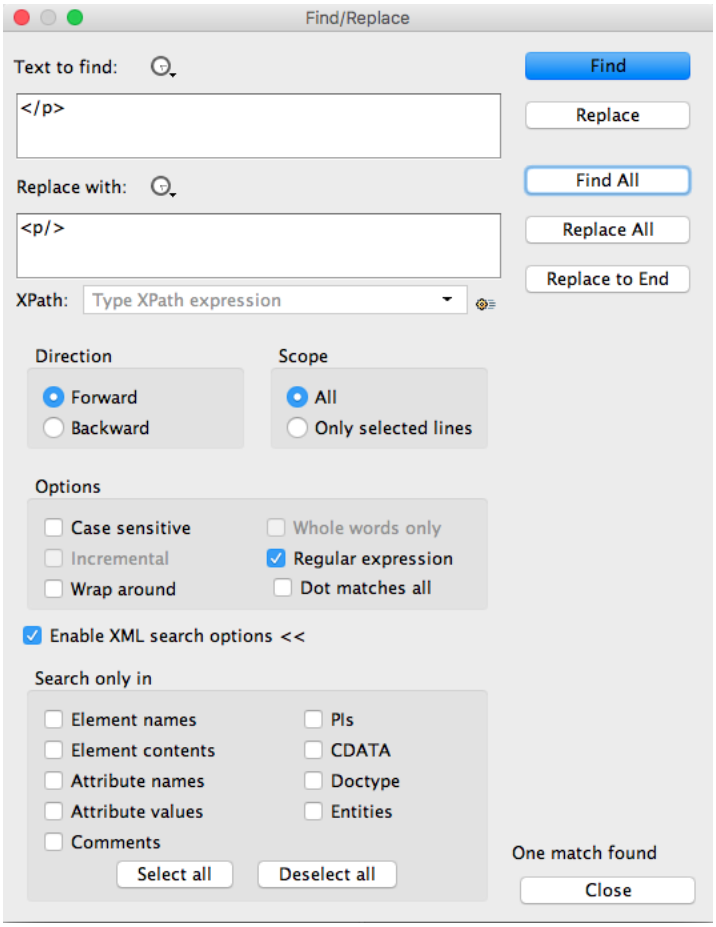


        ii. To remove <milestone> ELEMENTS with its various ATTRIBUTES, use the following find replace command:

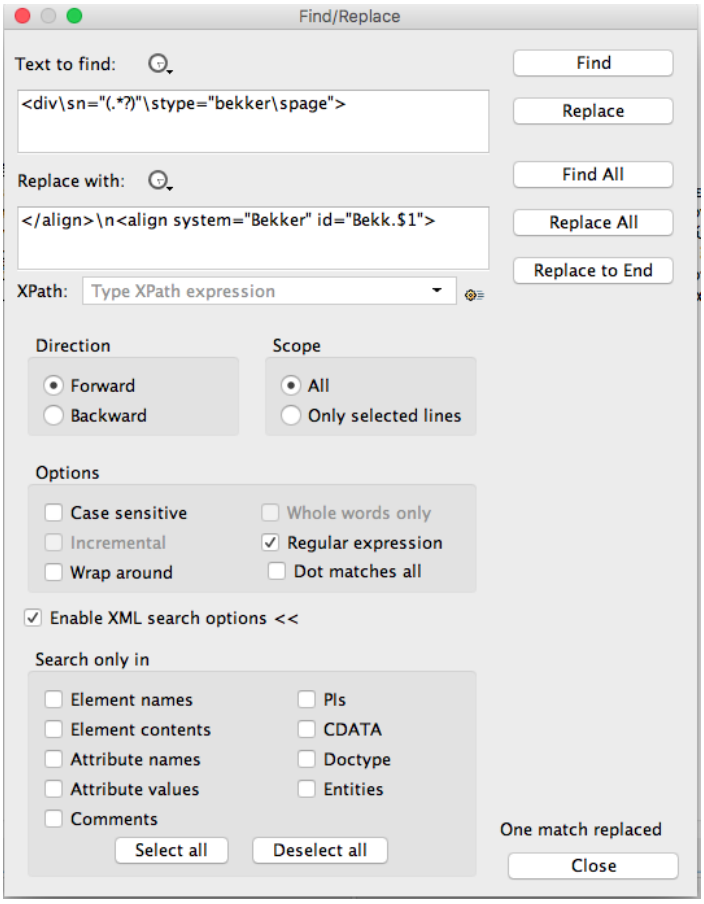


b. Replaced—In the above segment, the ELEMENT is not allowed by the syntax prescribed by the GoK text DTD. It preserves, however, valuable information (the Bekker number) as an ATTRIBUTE, which we want to include in the <align> ELEMENT in our DTD. We must, therefore, transform the ELEMENT into an <align> element. Likewise, we can modify the closing annotation so that it is null as required by the GoK text DTD.

    i. to replace with , use the following find and replace command:



ii. to transform the element into the <align> , use the following find and replace command:



(For more information on why this regular expression is used, see the instructions on using Regular Expressions at http://genealogiesofknowledge.net/genealogies-knowledge-corpus/#instructions)

In the find box: \s = whitespace

"(.*?)" tells the text editor to "grab" all the characters between the double quotes.

In the replace box: </align> is used to automatically close the previous open <align> ELEMENT.

Bekk.$1 tells the text editor to place the "grabbed" Bekker number immediately after the string Bekk.

Finally, the DTD requires an Alphabetic character and numeral. Thus, the following syntax should be used in @id in the <align< ELEMENT.

For **Aristotle** @id should have the characters Bekk. (NB: including the period) followed by the Bekker number. Thus, for example:

<align system="Bekker" id="Bekk.1095b">

For **Plato**, @id should have the characters Steph. (NB: including the period) followed by the Stephanus number. Thus, for example:

<align system="Stephanus" id="Steph.98c">

For **Galen**, @id should have the characters volume number of the text in Kühn's edition **in Roman Numerals** + . + Kühn page number. Thus, for example:

<align system="Kühn" id="XVIIa.657">

For **Hippocrates**, @id should have the volume number of the text in Littré's edition + . + Littré's page number. Thus, for example:

<align system="Littré" id="V.128">

The following text is what remains after making these changes:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE goktext SYSTEM "goktext.dtd">
<goktext>
   <section id="s1">
   <align system="Bekker" id="Bekk.1094a">
      πᾶσα τέχνη καὶ πᾶσα μέθοδος, ὁμοίως δὲ πρᾶξίς τε καὶ προαίρεσις, ἀγαθοῦ τινὸς
ἐφίεσθαι δοκεῖ· διὸ καλῶς ἀπεφήναντο τἀγαθόν, οὗ πάντ' ἐφίεται. διαφορὰ δέ τις φαίνεται
τῶν τελῶν· τὰ μὲν γάρ εἰσιν ἐνέργειαι, τὰ δὲ παρ' αὐτὰς ἔργα τινά. ὧν δ' εἰσὶ τέλη τινὰ παρὰ
τὰς πράξεις, ἐν τούτοις βελτίω πέφυκε τῶν ἐνεργειῶν τὰ ἔργα. πολλῶν δὲ πράξεων οὐσῶν
καὶ τεχνῶν καὶ ἐπιστημῶν πολλὰ γίνεται καὶ τὰ τέλη· ἰατρικῆς μὲν γὰρ ὑγίεια, ναυπηγικῆς
δὲ πλοῖον, στρατηγικῆς δὲ νίκη, οἰκονομικῆς δὲ πλοῦτος. ὅσαι δ' εἰσὶ τῶν τοιούτων ὑπὸ
μίαν τινὰ δύναμιν, καθάπερ ὑπὸ τὴν ἱππικὴν χαλινοποιικὴ καὶ ὅσαι ἄλλαι τῶν ἱππικῶν
ὀργάνων εἰσίν, αὕτη δὲ καὶ πᾶσα πολεμικὴ πρᾶξις ὑπὸ τὴν στρατηγικήν, κατὰ τὸν αὐτὸν δὴ
τρόπον ἄλλαι ὑφ' ἑτέρας· ἐν ἁπάσαις δὲ τὰ τῶν ἀρχιτεκτονικῶν τέλη πάντων ἐστὶν
αἱρετώτερα τῶν ὑπ' αὐτά· τούτων γὰρ χάριν κἀκεῖνα διώκεται. διαφέρει δ' οὐδὲν τὰς
ἐνεργείας αὐτὰς εἶναι τὰ τέλη τῶν πράξεων ἢ παρὰ ταύτας ἄλλο τι, καθάπερ ἐπὶ τῶν
λεχθεισῶν ἐπιστημῶν. εἰ δή τι τέλος ἐστὶ τῶν πρακτῶν ὃ δι' αὑτὸ βουλόμεθα, τἆλλα δὲ διὰ
τοῦτο, καὶ μὴ πάντα δι' ἕτερον αἱρούμεθα 'πρόεισι γὰρ οὕτω γ' εἰς ἄπειρον, ὥστ' εἶναι
κενὴν καὶ ματαίαν τὴν ὄρεξιν', δῆλον ὡς τοῦτ' ἂν εἴη τἀγαθὸν καὶ τὸ ἄριστον. ἆρ' οὖν καὶ
πρὸς τὸν βίον ἡ γνῶσις αὐτοῦ μεγάλην ἔχει ῥοπήν, καὶ καθάπερ τοξόται σκοπὸν ἔχοντες
μᾶλλον ἂν τυγχάνοιμεν τοῦ δέοντος; εἰ δ' οὕτω, πειρατέον τύπῳ γε περιλαβεῖν αὐτό τί ποτ'
ἐστὶ καὶ τίνος τῶν ἐπιστημῶν ἢ δυνάμεων. δόξειε δ' ἂν τῆς κυριωτάτης καὶ μάλιστα
ἀρχιτεκτονικῆς. τοιαύτη δ' ἡ πολιτικὴ φαίνεται· τίνας γὰρ εἶναι χρεὼν τῶν ἐπιστημῶν ἐν ταῖς
πόλεσι,<p/></align>
   </section>
</goktext>
```

Please note that the final 'cleaned' text **should not contain any non-essential white space** (blank lines, tabbed spaces etc.). There should not be more than a single line break between each new paragraph and all tabbed spaces should also be removed. Once again, white space can generally be removed using Regular Expressions. Please see the following page for more details: http://genealogiesofknowledge.net/genealogies-knowledge-corpus/#instructions