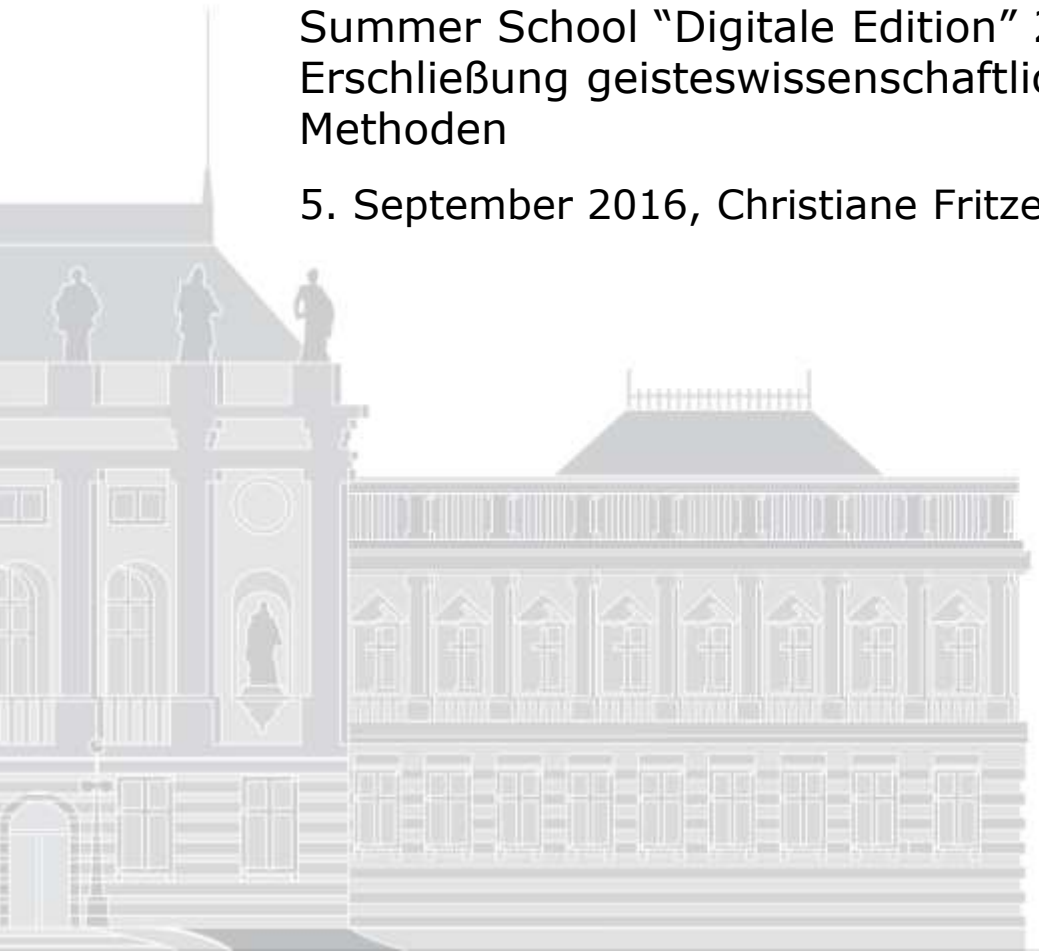
Textkodierung mit XML

Summer School "Digitale Edition" 2016
Erschließung geisteswissenschaftlicher Quellen mit digitalen Methoden

5. September 2016, Christiane Fritze



Zentrum für Informationsmodellierung
Austrian Centre for Digital Humanities
Elisabethstraße 59/III, SR 81.31

Überblick

- XML-Grundlagen: Was ist XML und wie geht das?
- Das XML-Dokument und seine Regeln
- XML schreiben - erste Fingerübung
- XML – und wie weiter?
- Zusammenfassung

XML Grundlagen

- XML heißt e**X**tensible **M**arkup **L**anguage. → erweiterbar
- XML ist ein internationaler W3C-Standard.
  - 10.2.1998 erste Empfehlung
  - 16.8.2006: 4. Edition ist der aktuell gültige Standard XML 1.1.

Was ist XML?

- … ein weit verbreiteter Standard für die **Beschreibung** und den **Austausch** von Daten.
- … **trennt Struktur und Darstellung** voneinander, Inhalt und Präsentation.
- … XML-Dokumente werden nach einem **Dokumentenmodell** entwickelt.
  - Dokumentenmodell als XML-Schema
- … **menschen- und maschinenlesbar**

```
<element attributname="attribut">
    Mein Text steht hier.
</element>
```

Was ist XML?

- XML ist eine universelle Metasprache.
- Einzelne Vokabulare
  - Z.B. xHTML, TEI, EAD, METS/MODS, SVG etc.
- Verschiedene Nutzungen, z.B.
  - Webpräsentation
  - Visualisierungen
  - Druckausgaben
- Menschen- und maschinenlesbar
- einfach

```
<element attributname="attribut">
    Mein Text steht hier.
</element>
```

Was ist XML?

- XML wird von einer breiten Softwarepalette unterstützt.
- XML hat eine große internationale Nutzer- und Entwickler-gemeinde.
- XML ist die Grundlage vieler Anwendungsstandards („XML is everywhere“).
- XML umfasst eine ganze Familie von begleitenden Standards.

```
<element attributname="attribut">
    Mein Text steht hier.
</element>
```

Überblick

- XML-Grundlagen: Was ist XML und wie geht das?
- Das XML-Dokument und seine Regeln
- XML schreiben - erste Fingerübung.
- XML – und wie weiter?
- Zusammenfassung

XML-Regeln

- Alles ist Text, Zeichendaten
- Markup bedeutet, Text mit Auszeichnungen zu versehen ...
- Dafür gibt es (Auszeichnungs-)Elemente.
  - Öffnendes Tag ... Text ... Schließendes Tag

```
<elementname>Elementinhalt</elementname>
```

- Elemente können Attribute haben.

```
<elementname
attributname="attributwert“>Elementinhalt</ele
mentname>
```

- Elemente enthalten Elemente, Text, beides - oder nichts.

```
<WasSollDas></WasSollDas> = <WasSollDas/>
```

XML-Regeln

- Alle Elemente müssen richtig geschachtelt sein.
- Es gibt ein – und nur ein - Wurzelelement.

XML-Regeln: Elemente

- Elementnamen müssen mit einem **Buchstaben, Unterstrich** oder **Doppelpunkt beginnen.**
  - Ergo: Sie dürfen nicht mit Zahlen beginnen.
- Elementnamen unterscheiden Groß- und Kleinschreibung.
  - `<Postkarte>` ≠ `<postkarte>`
- Elementnamen können Buchstaben, Zahlen, Bindestriche, Punkte oder Unterstriche, Umlaute und Akzente enthalten.
- Elementnamen dürfen nicht mit `xml` beginnen.
- Elementnamen können beliebig lang sein.
- Die Verwendung von `<` `>` `&` `&apos;` und `&quot;` ist nicht erlaubt.

XML-Regeln: Attribute

- Ein Element darf beliebig viele Attribute tragen.
- Ein Element darf nicht zweimal das gleiche Attribut haben.
- Attributwerte müssen in Anführungszeichen stehen.

```
color="red"
```

- Attributwerte dürfen alle möglichen Zeichen enthalten

XML-Regeln: Dokumentaufbau

- XML-Deklaration

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

- Processing instructions

```
<?xml-stylesheet type="text/xsl"
href=„transformieren.xsl"?>
```

- „eigentlichen“ Elemente, schön geschachtelt

```
<Postkarte><Anrede>Lieber
<Name>Paul</Name></Anrede>
</Postkarte>
```

- Kommentare

```
<!- hier steht ein Kommentar -->
```

XML-Regeln: Entitäten

- Was ist mit bereits durch XML besetzen Zeichen?
  - Dafür gibt es spezielle Entitäten:

    < wird zu `<`

    > wird zu `>`

    & wird zu `&`

    “ wird zu `&quot;`

    ‘ wird zu `&apos;`

Alternativ können sie als CDATA-Sektionen markiert werden:

```
<![CDATA[  Inhalte mit <spitzen>
Klammern ]]>
```

XML-Prüfung

- Die Prüfung erledigt ein Parser.
- Prüfung auf Wohlgeformtheit.
  - Ein XML-Dokument ist **wohlgeformt**, wenn es den Regeln des XML-Standards entspricht.
- Prüfung auf Validität/Gültigkeit.
  - Ein XML-Dokument ist **valide**, wenn es wohlgeformt ist und der Grammatik des XML-Schemas entspricht.

Überblick

- XML-Grundlagen: Was ist XML und wie geht das?
- Das XML-Dokument und seine Regeln
- XML schreiben - erste Fingerübung.
- XML – und wie weiter?
- Zusammenfassung

XML schreiben: Hilfsmittel

- einfache Textprogramme (MS Editor, TextPad, BBedit)
- XML-Editoren
  - erleichtern die Arbeit mit XML, z.B.
  - XML Notepad http://www.microsoft.com/en-us/download/details.aspx?id=7973
  - XMLSpy http://www.altova.com/de/xmlspy.html
  - oXygen http://oxygenxml.com/
  - jEdit http://www.jedit.org/

XML selber schreiben

- *Ein* Beispiel: eine Postkarte

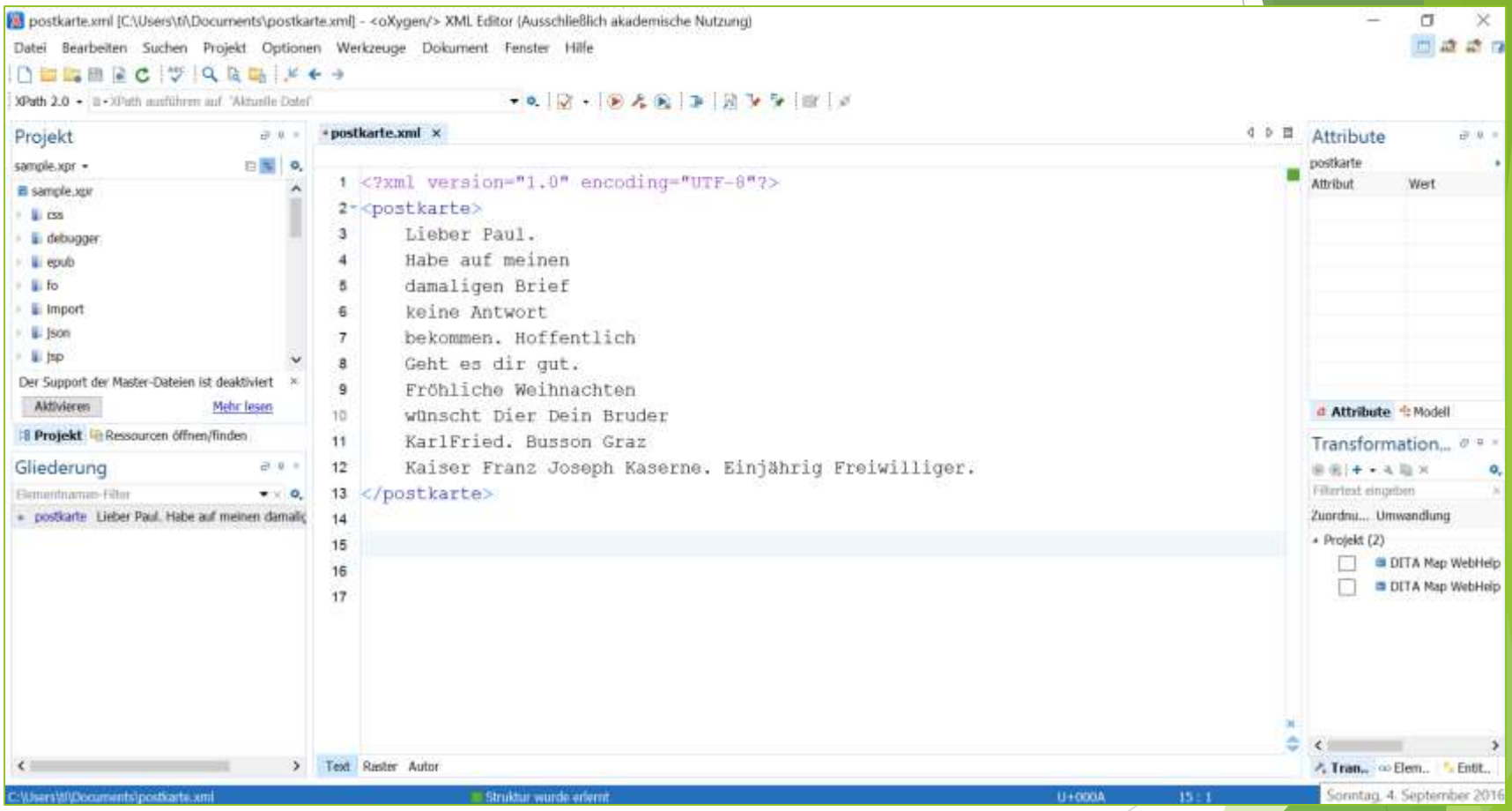
Graz, Stadtpfarrkirche und Herrengasse, Verlag Josef Kienreich, Graz, gelaufen von Graz nach Wien, ca. 1906–1908, mit Autotypie kombinierter Lichtdruck, Graz Museum

Rückseite Graz, Stadtpfarrkirche und Herrengasse

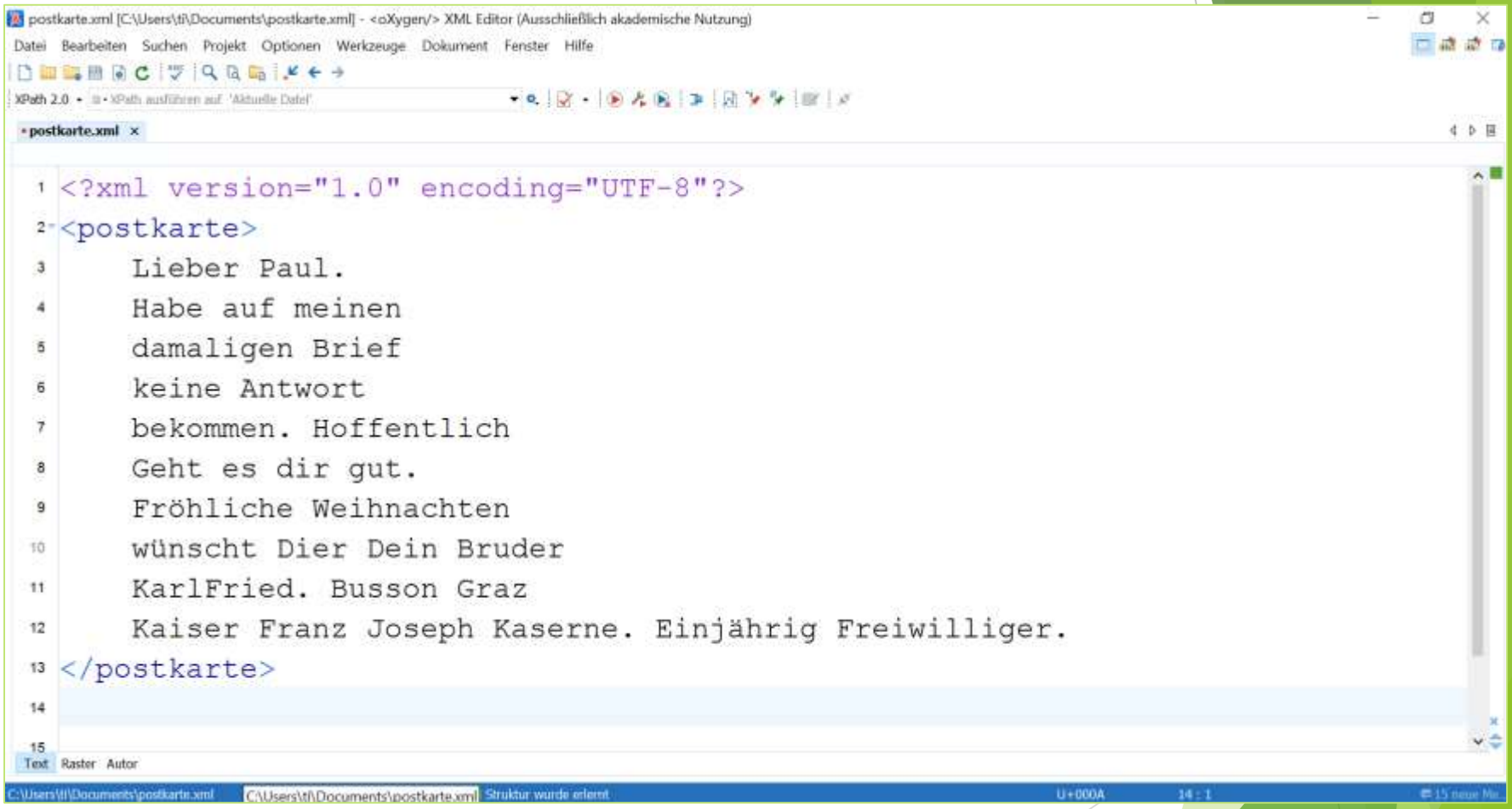
XML schreiben – im Editor

- Starten Sie das Programm Oxygen.
- Legen Sie eine neue XML-Datei an.
  - Strg+N | Menü Datei → Neue Datei
- Transkribieren Sie die Postkarte.
- Aufgabe:
  - Versuchen Sie, Ihr Wissen über das Dokument festzuhalten
  - Textbestand?
  - Strukturen?
  - Zusätzliches Wissen?
  - Welche Informationen werden gebraucht, um später eine Edition daraus zu machen?
- Lassen Sie sich helfen: rote Unterkringelungen?

XML selber schreiben



XML selber schreiben



Ihre Ideen zur Kodierung

- Was haben Sie ausgezeichnet?
- Element oder Attribut?

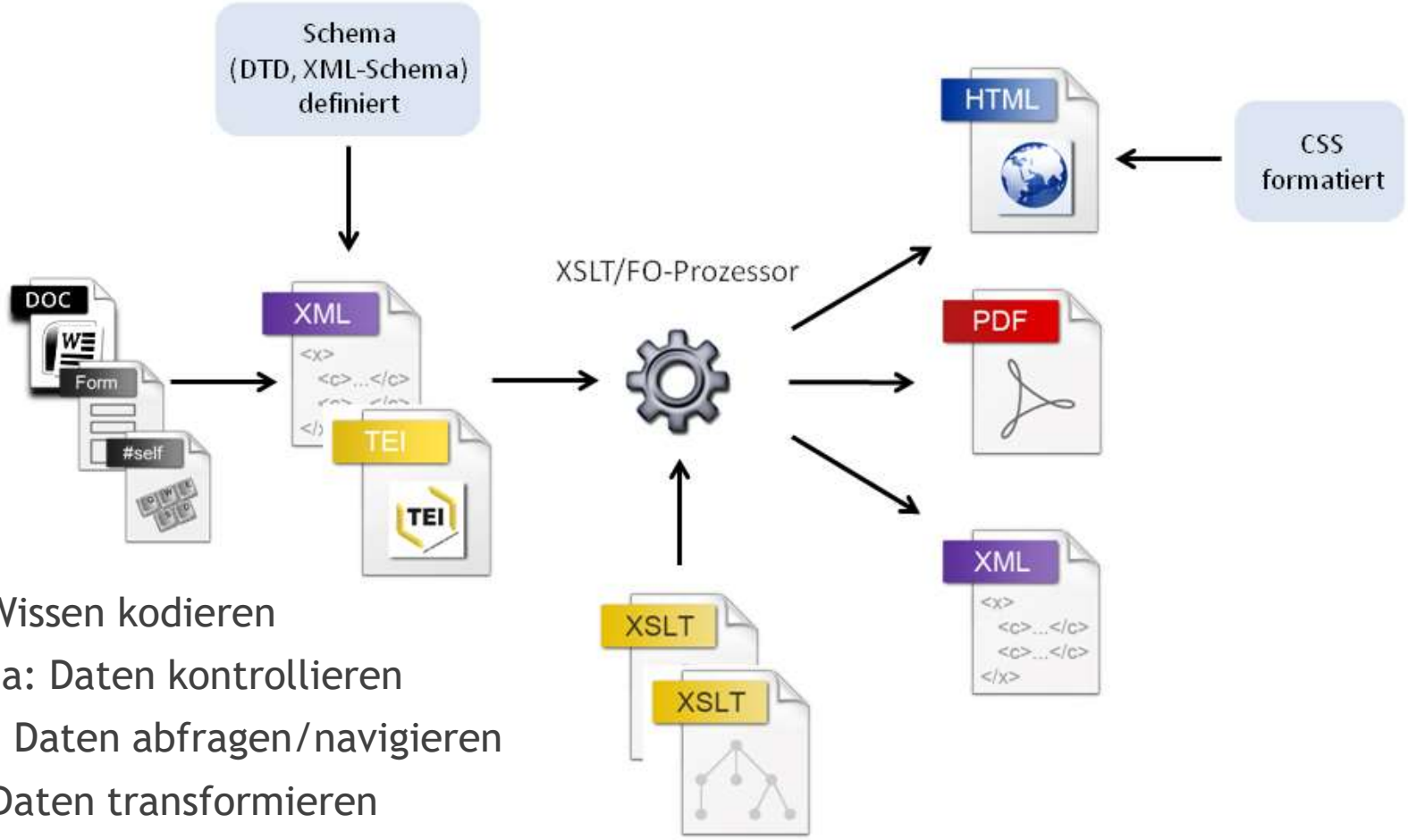
Überblick

- XML-Grundlagen: Was ist XML und wie geht das?
- Das XML-Dokument und seine Regeln
- XML schreiben - erste Fingerübung.
- XML – und wie weiter?
- Zusammenfassung

XML-Familie

- XML … beschreibt strukturierte Daten
- XPath … erlaubt die Navigation in XML-Daten
- XML Schema … beschreibt ein striktes Datenmodell
- XSL … eXtensible Style Language
  - XSLT … transformiert XML-Dokumente
  - XSL-FO … beschreibt eine formatierte Ausgabe (z.B. für den Druck)
- XQuery … ist eine XML-Datenbankabfragesprache
- XInclude … lädt XML-Dokumente ineinander
- XLink … beschreibt komplexe Links
- XPointer … ermöglich komplexe Referenzierung
- XForms … beschreibt Eingabeformulare

XML-basierte Dokumentverarbeitung



XML: Wissen kodieren

Schema: Daten kontrollieren

XPath: Daten abfragen/navigieren

XSLT: Daten transformieren

Überblick

- XML-Grundlagen: Was ist XML und wie geht das?
- Das XML-Dokument und seine Regeln
- XML schreiben - erste Fingerübung.
- XML – und wie weiter?
- Zusammenfassung

Kurz: XML ist…

- XML ist einfach.
- XML ist wie HTML, nur anders:
  - XML ist erweiterbar.
  - XML muss wohlgeformt sein.
  - XML kann validiert werden.
- besonders gut geeignet, um Texte zu strukturieren.

Alles über ein XML-Dokument

- Ein XML-Dokument kann enthalten:
  - Elemente, eventuell mit Attributen
  - Verarbeitungsanweisungen
  - Kommentare
  - Entitätsreferenzen
- Ein XML-Dokument muss wohlgeformt sein und kann validiert werden.
- XML Attributwerte müssen in ” ” stehen.
- XML-Dokumente sind als lineare Zeichenketten kodiert.
- XML-Dokumente beginnen mit einer besonderen Verarbeitungsanweisung.

Anhang

- Literatur
- Gedanken zu XML
- Benutzung des Oxygen-Editors

Literatur zu XML

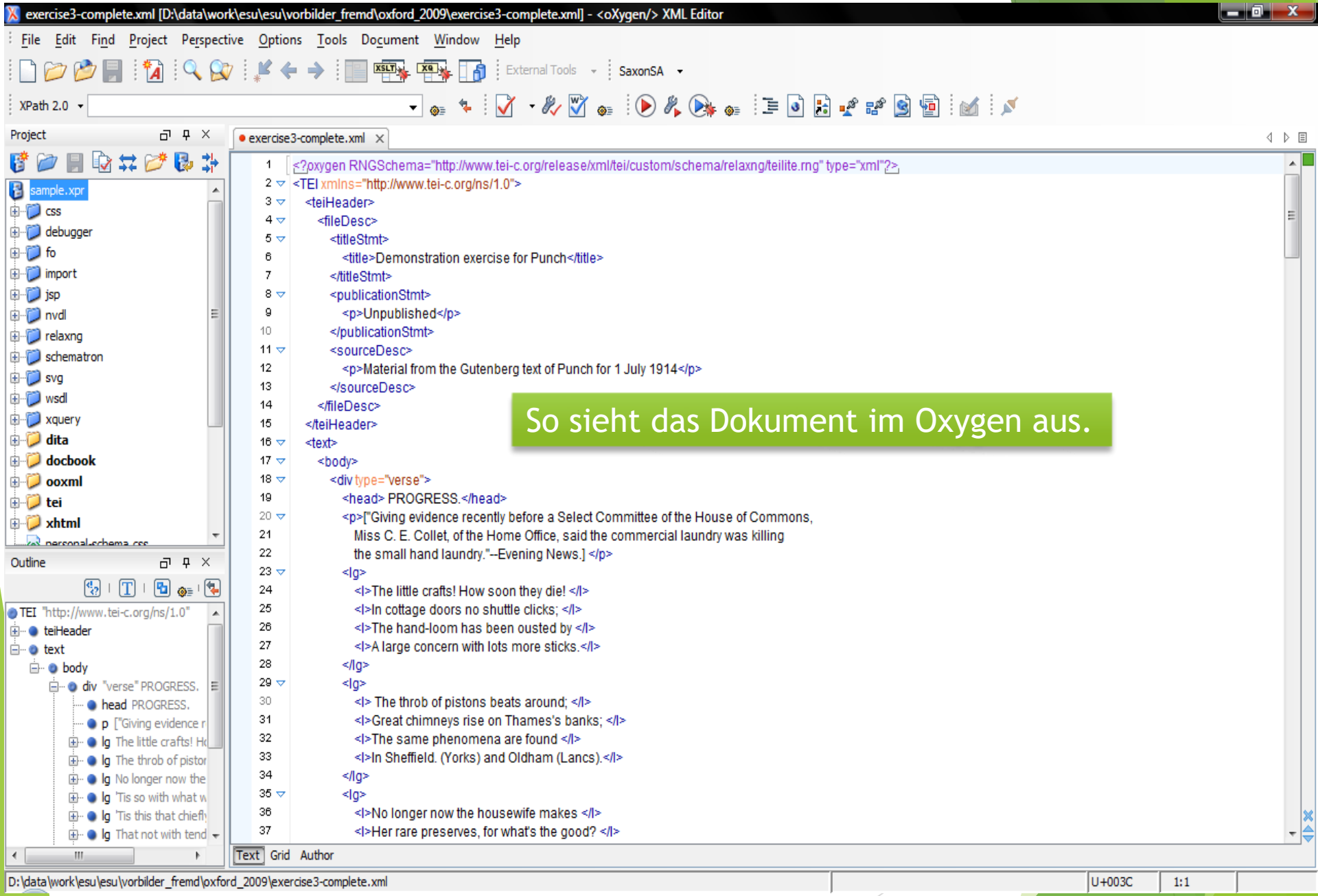
- Vonhoegen, Helmut, *Einstieg in XML. Grundlagen, Praxis, Referenz*, Galileo Press, Bonn 2009[5].
- St. Laurent, Simon/Fitzgerald, Michael, *XML. kurz & gut*, O'Reilly, Köln 2006[3].
- W3 Schools Tutorial: http://www.w3schools.com/xml/ (XML),
- W3C Spezifikation: http://www.w3.org/TR/xml/

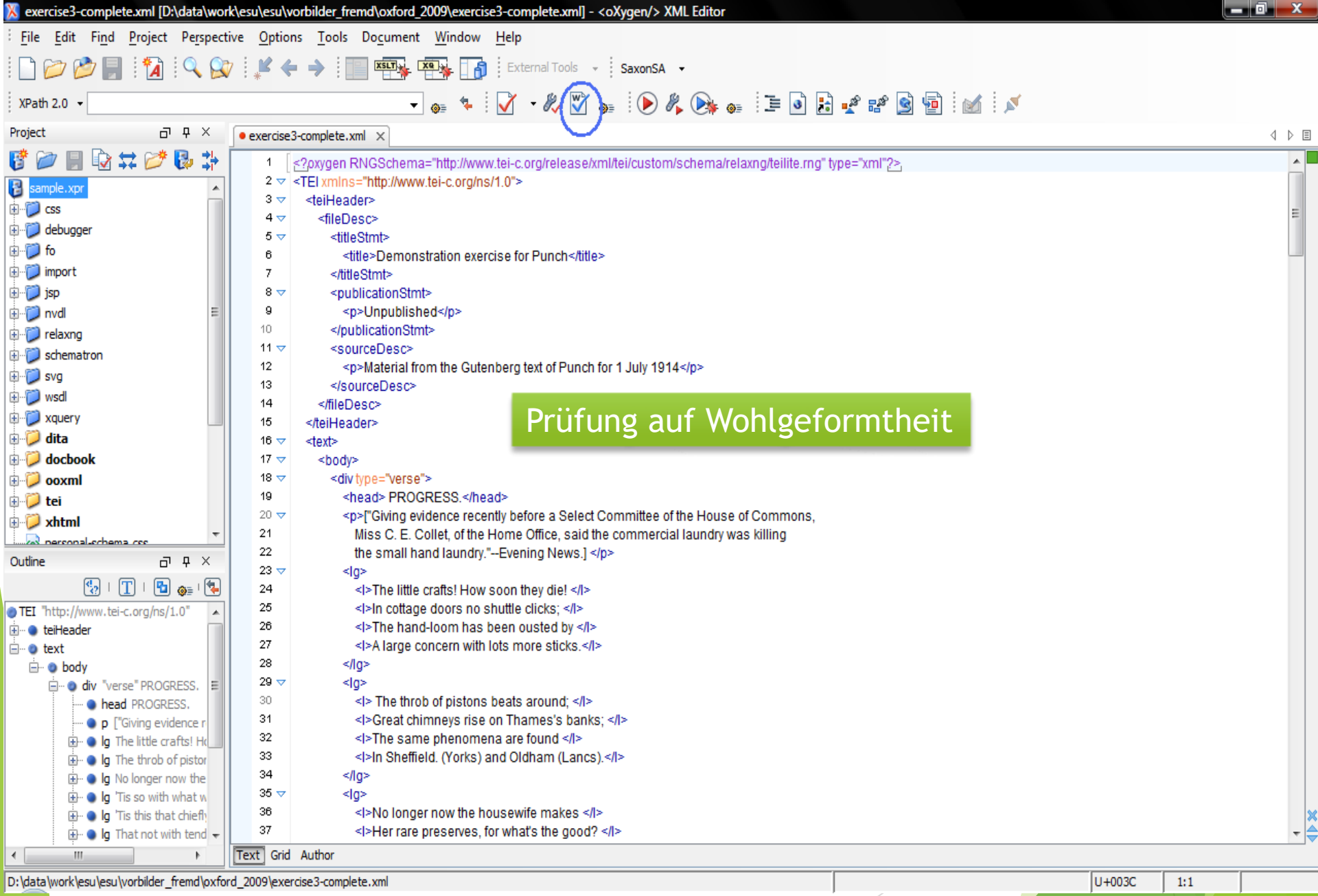
XML - was die Leute sagen…

- XML ist ein Akronym
- XML steht für eXtensible Markup Language
- XML ist "plain text"
- XML ist einfach
- XML ist wie HTML, nur anders
- XML ist eine Auszeichnungssprache
- XML sagt: "invent your own tags"
- XML ist eine Metasprache
- XML ist eine Grammatik
- XML ist eine Datenstruktur (ein Datenformat? Ein Datenmodell?)
- XML beschreibt strukturierte Informationen
- XML ist ein Baum
- XML ist ein serialisierter Graph
- XML ist eine hierarchisierte Sequenz
- XML ist ein (Daten-?)Standard
- XML ist ein Standard des W3C
- XML ist eine Untermenge von SGML
- XML ist plattformunabhängig
- XML unterstützt
- XML wird von einer breiten Softwarepalette ist zukunftssicher
- XML ist flexibel
- XML kann ein Austauschformat sein
- XML ist eine Notation
- XML ist dokumentnah
- XML ist die Grundlage von Online-Ressourcen
- xHTML ist XML
- XML ist mächtig, weil es abstrahiert und weil es viele "Freunde" hat
- XML "is everywhere”

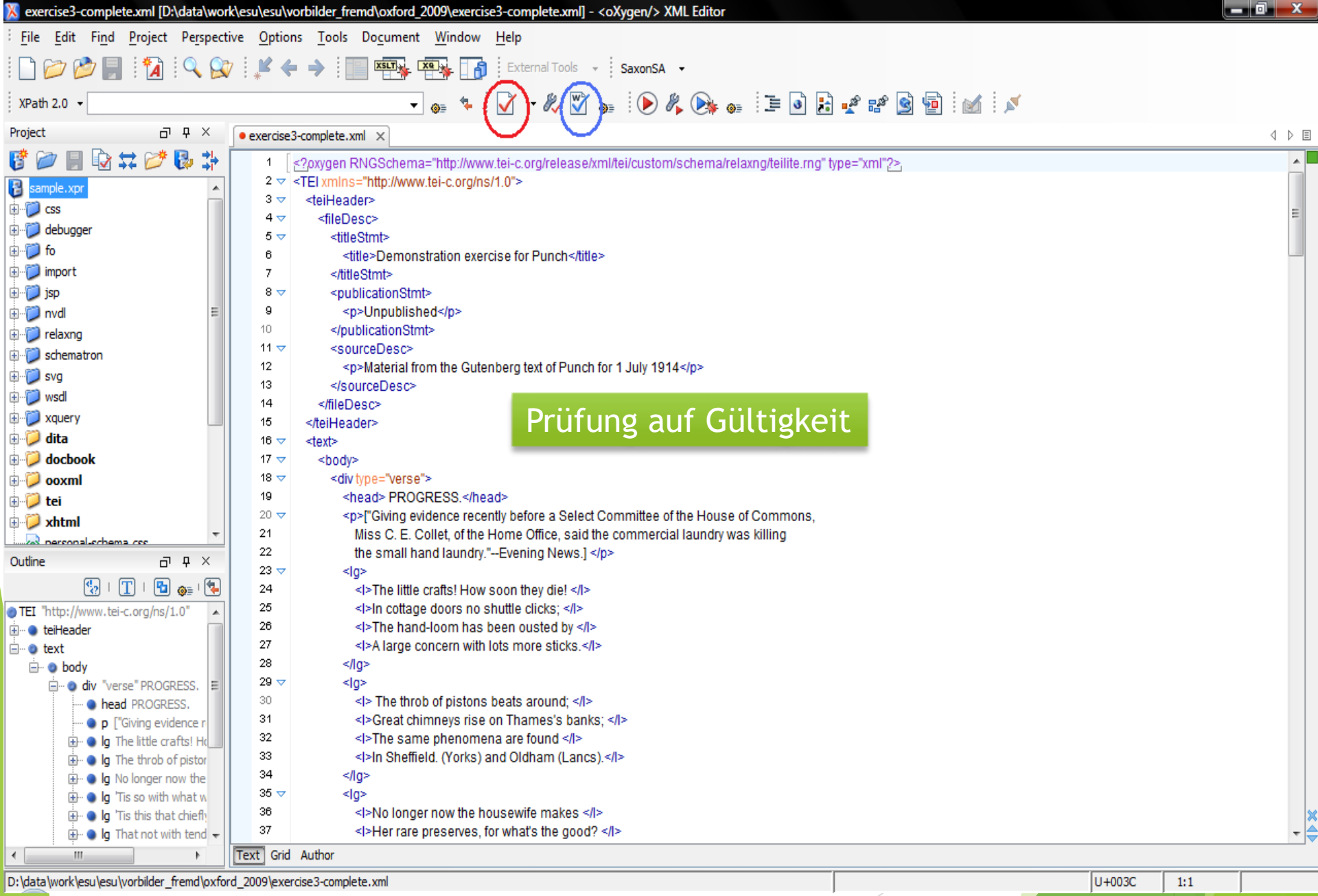
Warum Oxygen

- plattformunabhängig
- als Eclipse plug-in erhältlich
- subversion client für kollaboratives Arbeiten
- TEI Unterstützung
- Unterstützung aller Schemasprachen
- Syntaxvervollständigung
- Tool-tip Dokumentation
- XSLT und FOP Unterstützung für Transformationen nach HTML, PDF...

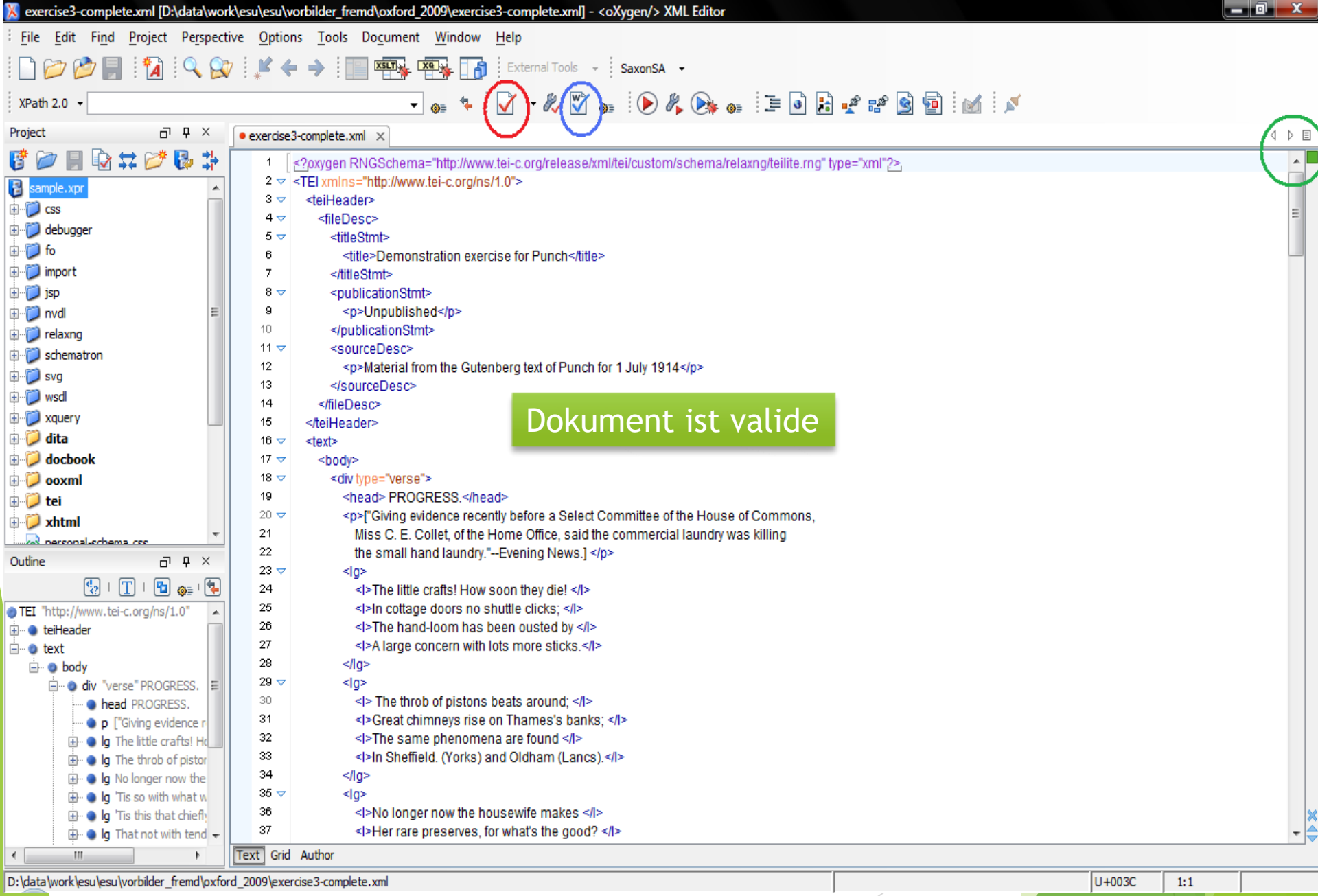

exercise3-complete.xml [D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml] - <oXygen/> XML Editor
File Edit Find Project Perspective Options Tools Document Window Help
External Tools
SaxonSA
XPath 2.0
Project
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
ooxml
tei
xhtml
Outline
TEI "http://www.tei-c.org/ns/1.0"
teiHeader
text
body
div "verse" PROGRESS.
head PROGRESS.
exercise3-complete.xml
1 <?oxygen RNGSchema="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/teilite.rng" type="xml"?>
2 <TEI xmlns="http://www.tei-c.org/ns/1.0">
3 <teiHeader>
4 <fileDesc>
5 <titleStmt>
6 <title>Demonstration exercise for Punch</title>
7 </titleStmt>
8 <publicationStmt>
9 Unpublished
10 </publicationStmt>
11 <sourceDesc>
12 Material from the Gutenberg text of Punch for 1 July 1914
13 </sourceDesc>
14 </fileDesc>
15 </teiHeader>
16 <text>
17 <body>
18
19 <head> PROGRESS.</head>
20 ["Giving evidence recently before a Select Committee of the House of Commons,
21 Miss C. E. Collet, of the Home Office, said the commercial laundry was killing
22 the small hand laundry."--Evening News.]
23 <lg>
24 <l>The little crafts! How soon they die! </l>
25 <l>In cottage doors no shuttle clicks; </l>
26 <l>The hand-loom has been ousted by </l>
27 <l>A large concern with lots more sticks.</l>
28 </lg>
29 <lg>
30 <l> The throb of pistons beats around; </l>
31 <l>Great chimneys rise on Thames's banks; </l>
32 <l>The same phenomena are found </l>
33 <l>In Sheffield. (Yorks) and Oldham (Lancs).</l>
34 </lg>
35 <lg>
36 <l>No longer now the housewife makes </l>
37 <l>Her rare preserves, for what's the good? </l>
So sieht das Dokument im Oxygen aus.
Text Grid Author
D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml
U+003C
1:1


exercise3-complete.xml [D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml] - <oXygen/> XML Editor
File Edit Find Project Perspective Options Tools Document Window Help
External Tools
SaxonSA
XPath 2.0
Project
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
ooxml
tei
xhtml
Outline
TEI "http://www.tei-c.org/ns/1.0"
teiHeader
text
body
div "verse" PROGRESS.
head PROGRESS.
exercise3-complete.xml
<?oxygen RNGSchema="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/teilite.rng" type="xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Demonstration exercise for Punch</title>
</titleStmt>
<publicationStmt>
Unpublished
</publicationStmt>
<sourceDesc>
Material from the Gutenberg text of Punch for 1 July 1914
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>

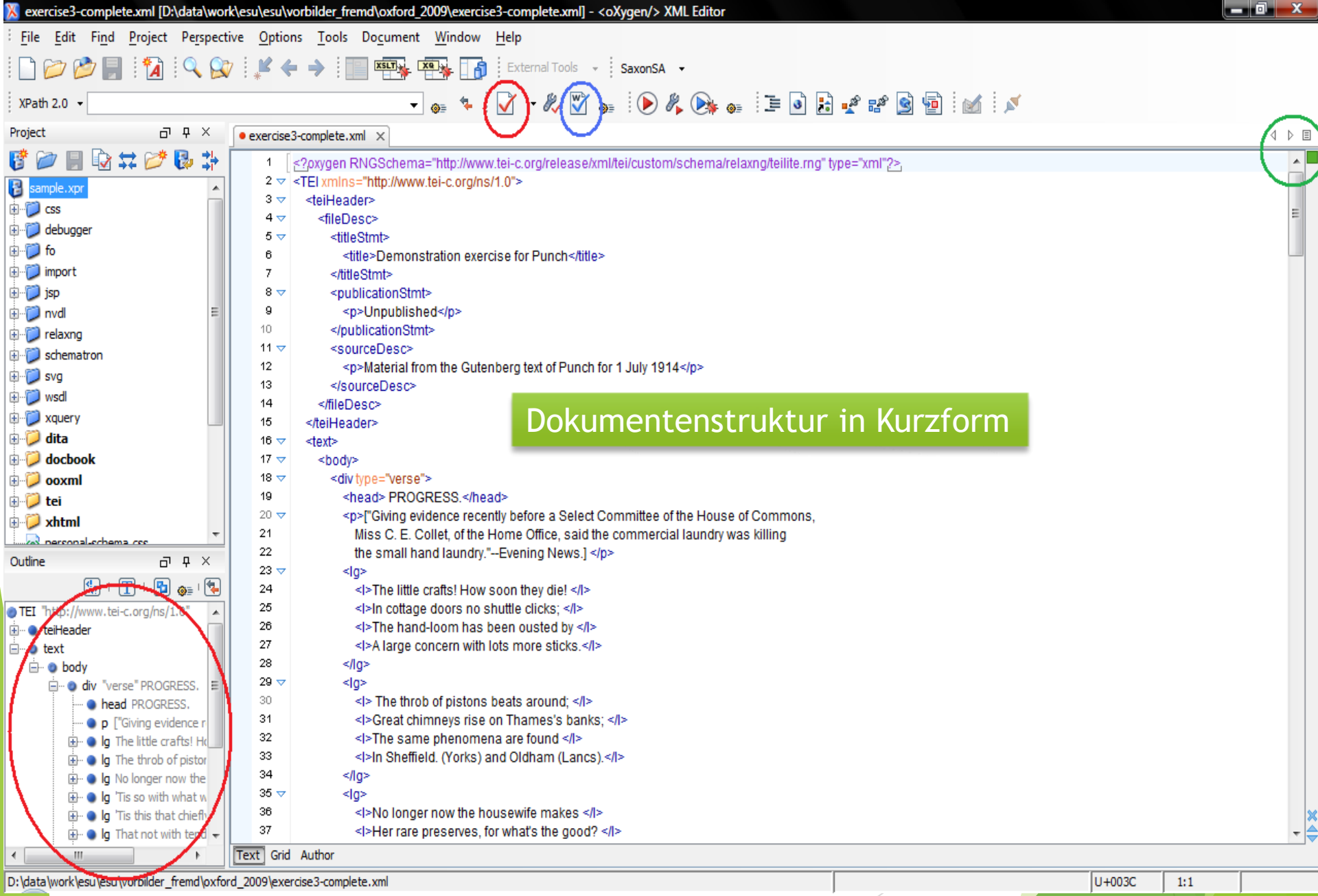
<head> PROGRESS.</head>
["Giving evidence recently before a Select Committee of the House of Commons,
Miss C. E. Collet, of the Home Office, said the commercial laundry was killing
the small hand laundry."--Evening News.]
<lg>
<l>The little crafts! How soon they die! </l>
<l>In cottage doors no shuttle clicks; </l>
<l>The hand-loom has been ousted by </l>
<l>A large concern with lots more sticks.</l>
</lg>
<lg>
<l> The throb of pistons beats around; </l>
<l>Great chimneys rise on Thames's banks; </l>
<l>The same phenomena are found </l>
<l>In Sheffield. (Yorks) and Oldham (Lancs).</l>
</lg>
<lg>
<l>No longer now the housewife makes </l>
<l>Her rare preserves, for what's the good? </l>
Prüfung auf Wohlgeformtheit
Text Grid Author
D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml
U+003C
1:1


exercise3-complete.xml [D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml] - <oXygen/> XML Editor
File Edit Find Project Perspective Options Tools Document Window Help
External Tools
SaxonSA
XPath 2.0
Project
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
ooxml
tei
xhtml
Outline
TEI "http://www.tei-c.org/ns/1.0"
teiHeader
text
body
div "verse" PROGRESS.
head PROGRESS.
exercise3-complete.xml
<?oxygen RNGSchema="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/teilite.rng" type="xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Demonstration exercise for Punch</title>
</titleStmt>
<publicationStmt>
Unpublished
</publicationStmt>
<sourceDesc>
Material from the Gutenberg text of Punch for 1 July 1914
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>

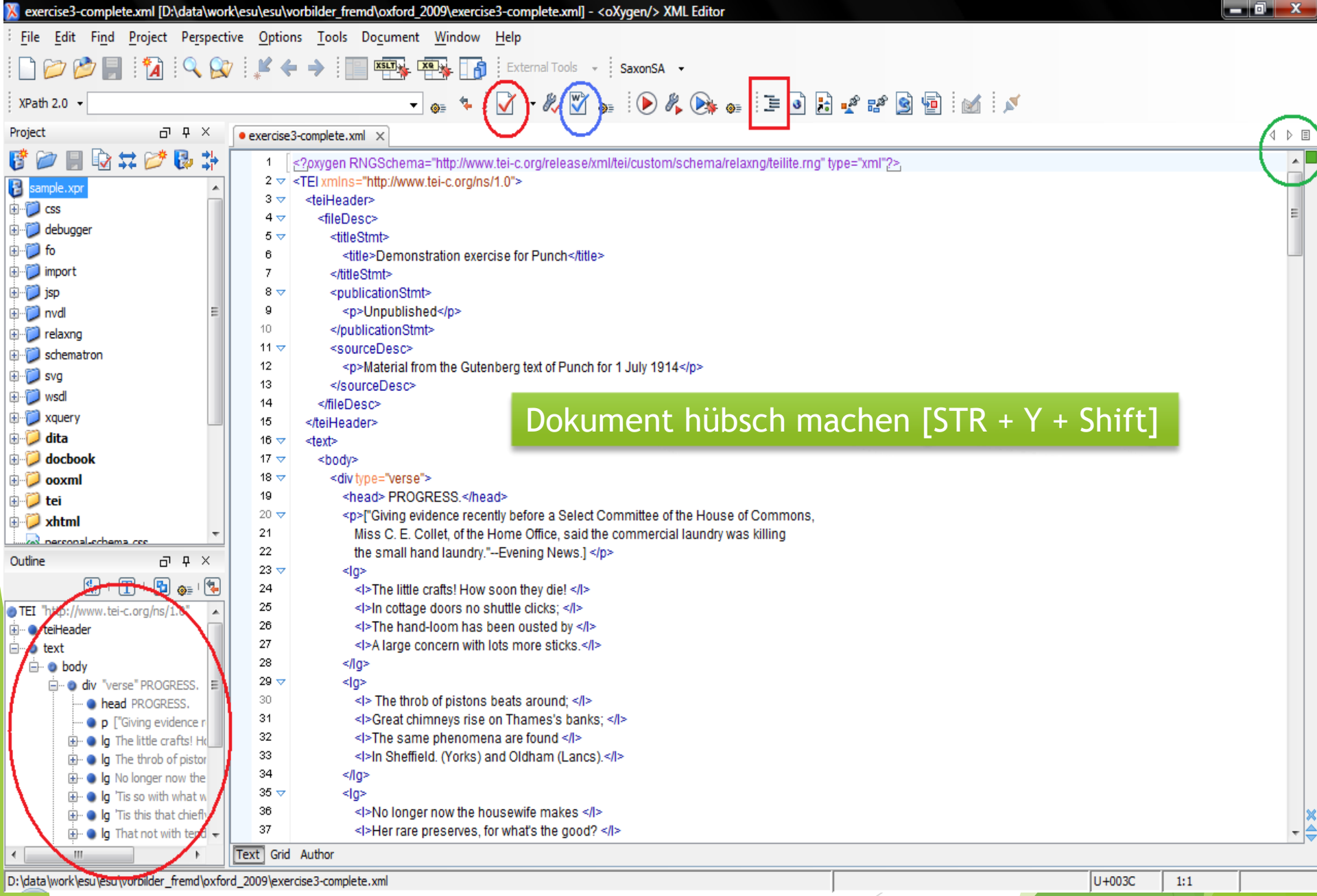
<head> PROGRESS.</head>
["Giving evidence recently before a Select Committee of the House of Commons,
Miss C. E. Collet, of the Home Office, said the commercial laundry was killing
the small hand laundry."--Evening News.]
<lg>
<l>The little crafts! How soon they die! </l>
<l>In cottage doors no shuttle clicks; </l>
<l>The hand-loom has been ousted by </l>
<l>A large concern with lots more sticks.</l>
</lg>
<lg>
<l> The throb of pistons beats around; </l>
<l>Great chimneys rise on Thames's banks; </l>
<l>The same phenomena are found </l>
<l>In Sheffield. (Yorks) and Oldham (Lancs).</l>
</lg>
<lg>
<l>No longer now the housewife makes </l>
<l>Her rare preserves, for what's the good? </l>
Prüfung auf Gültigkeit
Text Grid Author
D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml
U+003C
1:1


exercise3-complete.xml [D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml] - <oXygen/> XML Editor
File Edit Find Project Perspective Options Tools Document Window Help
External Tools
SaxonSA
XPath 2.0
Project
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
ooxml
tei
xhtml
Outline
TEI "http://www.tei-c.org/ns/1.0"
teiHeader
text
body
div "verse" PROGRESS.
head PROGRESS.
exercise3-complete.xml
<?oxygen RNGSchema="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/teilite.rng" type="xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Demonstration exercise for Punch</title>
</titleStmt>
<publicationStmt>
Unpublished
</publicationStmt>
<sourceDesc>
Material from the Gutenberg text of Punch for 1 July 1914
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>

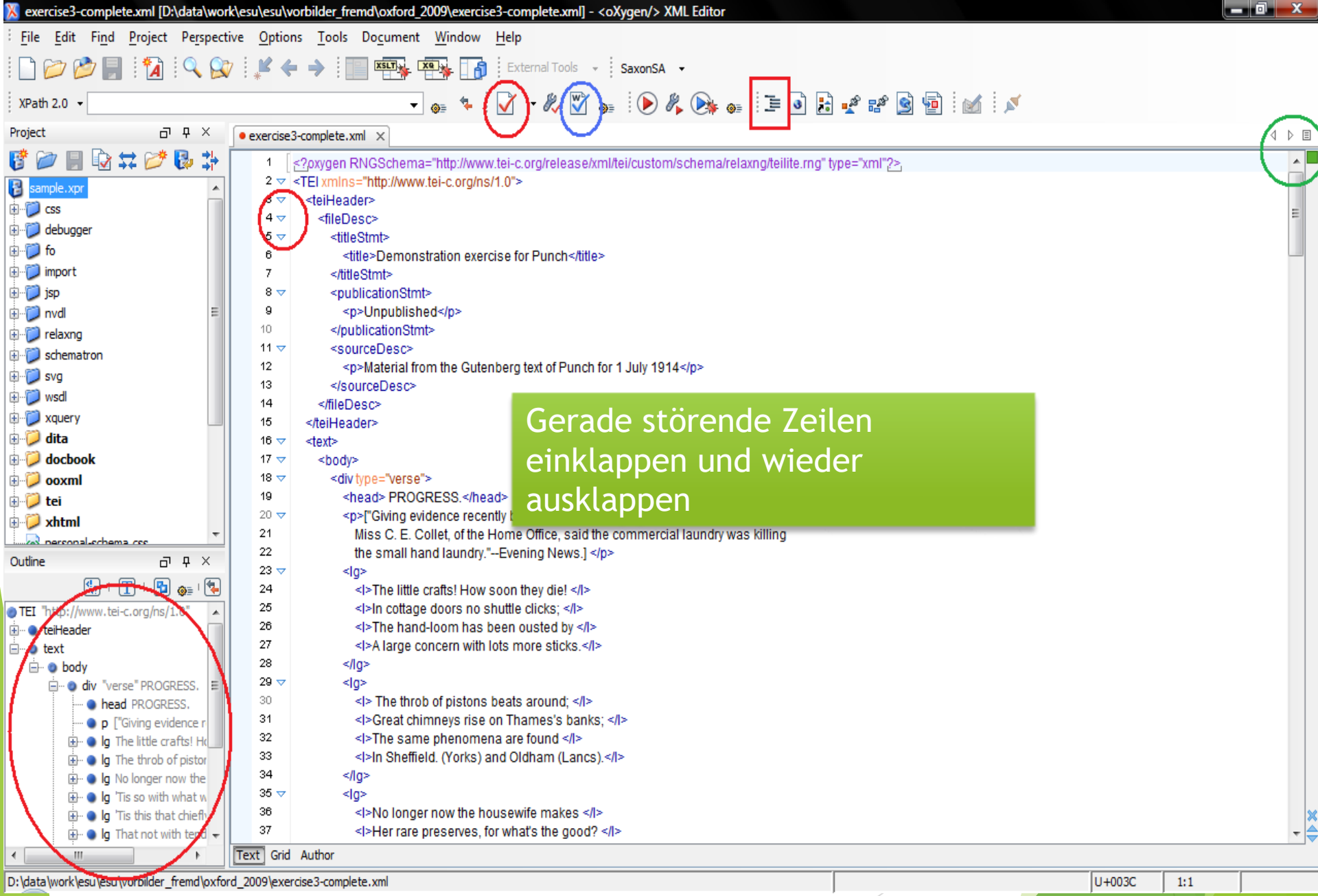
<head> PROGRESS.</head>
["Giving evidence recently before a Select Committee of the House of Commons,
Miss C. E. Collet, of the Home Office, said the commercial laundry was killing
the small hand laundry."--Evening News.]
<lg>
<l>The little crafts! How soon they die! </l>
<l>In cottage doors no shuttle clicks; </l>
<l>The hand-loom has been ousted by </l>
<l>A large concern with lots more sticks.</l>
</lg>
<lg>
<l> The throb of pistons beats around; </l>
<l>Great chimneys rise on Thames's banks; </l>
<l>The same phenomena are found </l>
<l>In Sheffield. (Yorks) and Oldham (Lancs).</l>
</lg>
<lg>
<l>No longer now the housewife makes </l>
<l>Her rare preserves, for what's the good? </l>
Dokument ist valide
Text Grid Author
D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml
U+003C
1:1


exercise3-complete.xml [D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml] - <oXygen/> XML Editor
File Edit Find Project Perspective Options Tools Document Window Help
External Tools
SaxonSA
XPath 2.0
Project
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
ooxml
tei
xhtml
Outline
TEI "http://www.tei-c.org/ns/1.0"
teiHeader
text
body
div "verse" PROGRESS.
head PROGRESS.
exercise3-complete.xml
<?oxygen RNGSchema="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/teilite.rng" type="xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Demonstration exercise for Punch</title>
</titleStmt>
<publicationStmt>
Unpublished
</publicationStmt>
<sourceDesc>
Material from the Gutenberg text of Punch for 1 July 1914
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>

<head> PROGRESS.</head>
["Giving evidence recently before a Select Committee of the House of Commons,
Miss C. E. Collet, of the Home Office, said the commercial laundry was killing
the small hand laundry."--Evening News.]
<lg>
<l>The little crafts! How soon they die! </l>
<l>In cottage doors no shuttle clicks; </l>
<l>The hand-loom has been ousted by </l>
<l>A large concern with lots more sticks.</l>
</lg>
<lg>
<l> The throb of pistons beats around; </l>
<l>Great chimneys rise on Thames's banks; </l>
<l>The same phenomena are found </l>
<l>In Sheffield. (Yorks) and Oldham (Lancs).</l>
</lg>
<lg>
<l>No longer now the housewife makes </l>
<l>Her rare preserves, for what's the good? </l>
Dokumentenstruktur in Kurzform
Text Grid Author
D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml
U+003C
1:1


exercise3-complete.xml [D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml] - <oXygen/> XML Editor
File Edit Find Project Perspective Options Tools Document Window Help
External Tools
SaxonSA
XPath 2.0
Project
sample.xpr
css
debugger
fo
import
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
docbook
ooxml
tei
xhtml
Outline
TEI
teiHeader
text
body
exercise3-complete.xml
<?oxygen RNGSchema="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/teilite.rng" type="xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Demonstration exercise for Punch</title>
</titleStmt>
<publicationStmt>
Unpublished
</publicationStmt>
<sourceDesc>
Material from the Gutenberg text of Punch for 1 July 1914
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>

<head> PROGRESS.</head>
["Giving evidence recently before a Select Committee of the House of Commons,
Miss C. E. Collet, of the Home Office, said the commercial laundry was killing
the small hand laundry."--Evening News.]
<lg>
<l>The little crafts! How soon they die! </l>
<l>In cottage doors no shuttle clicks; </l>
<l>The hand-loom has been ousted by </l>
<l>A large concern with lots more sticks.</l>
</lg>
<lg>
<l> The throb of pistons beats around; </l>
<l>Great chimneys rise on Thames's banks; </l>
<l>The same phenomena are found </l>
<l>In Sheffield. (Yorks) and Oldham (Lancs).</l>
</lg>
<lg>
<l>No longer now the housewife makes </l>
<l>Her rare preserves, for what's the good? </l>
Dokument hübsch machen [STR + Y + Shift]
Text Grid Author
D:\data\work\esu\esu\vorbilder_fremd\oxford_2009\exercise3-complete.xml
U+003C
1:1



Gerade störende Zeilen einklappen und wieder ausklappen