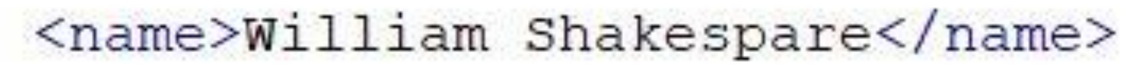
Digital Philology – ELTE-DH Winter School 2020

Zsófia Fellegi

Research Institute for Literary Studies

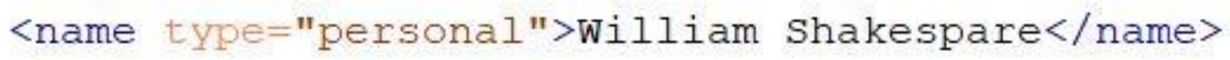
XML - memo

- Basic rules
  - close the tags
  - consistency
  - no overlaps



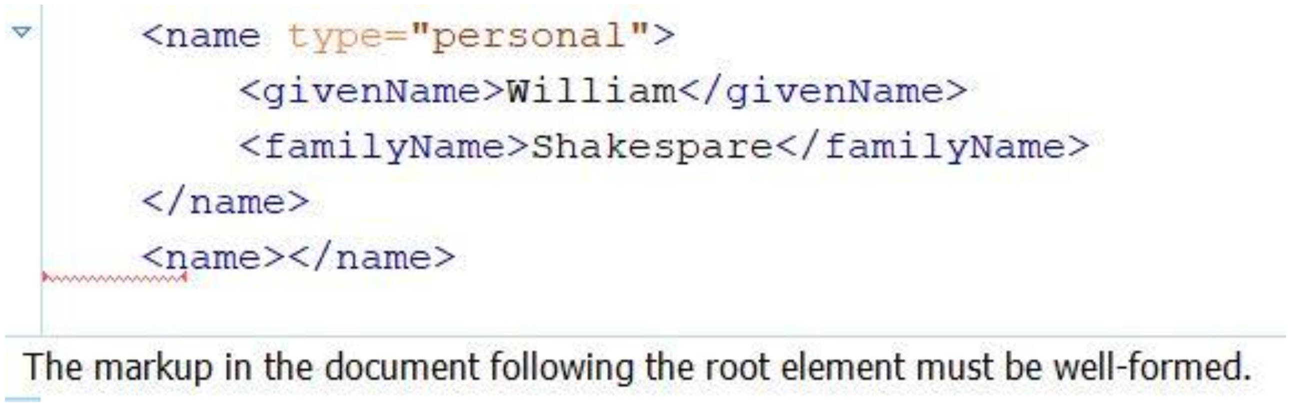
```
<name>William Shakespare</name>
```

- Attributes
  - additional information (eg. xml:id)



```
<name type="personal">William Shakespare</name>
```

- Root element



```
    <name type="personal">
        <givenName>William</givenName>
        <familyName>Shakespare</familyName>
    </name>
    <name></name>
```

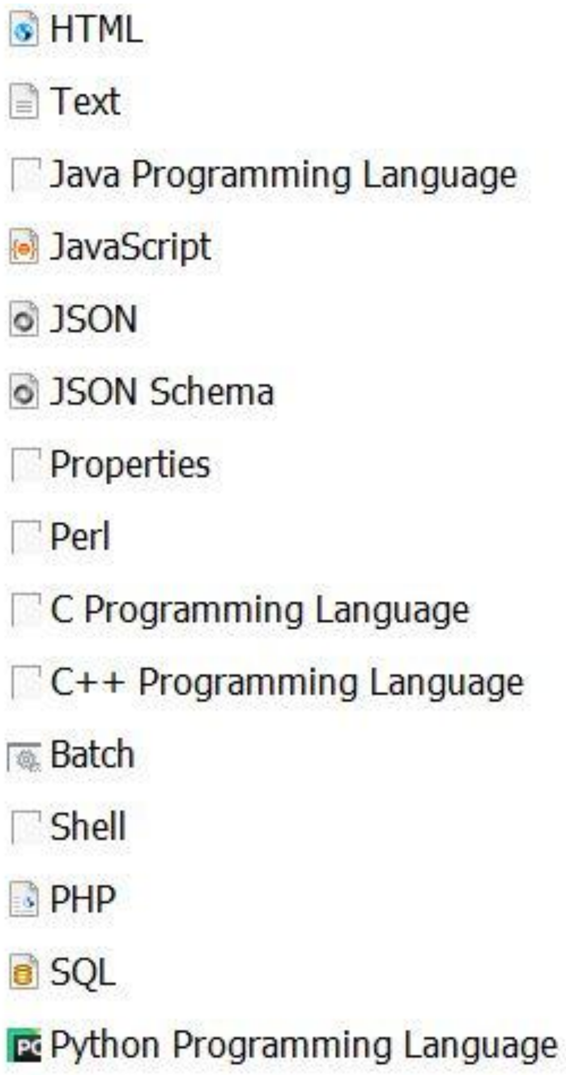
The markup in the document following the root element must be well-formed.

Text Encoding Initiative

- 1990 - TEI P1: Guidelines for the Encoding and Interchange of Machine-Readable Texts
- 2019. 07. 16 - TEI P5 3.6.0 □ 1934 p.
- TEI XML structure: header + text
- <teiHeader> : structured metadata
    - <fileDesc>,<encodingDesc>, <profileDesc>, <revisionDesc>
- <text> : structured text
    - <body>, <div>, <p>, <lg>

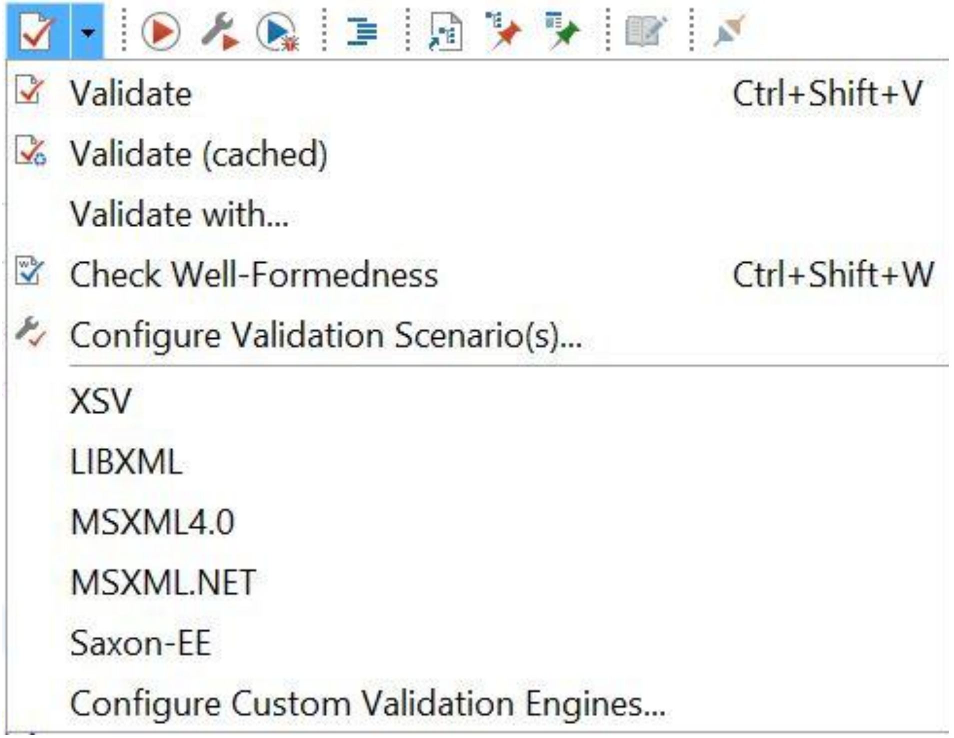
Oxygen XML Editor

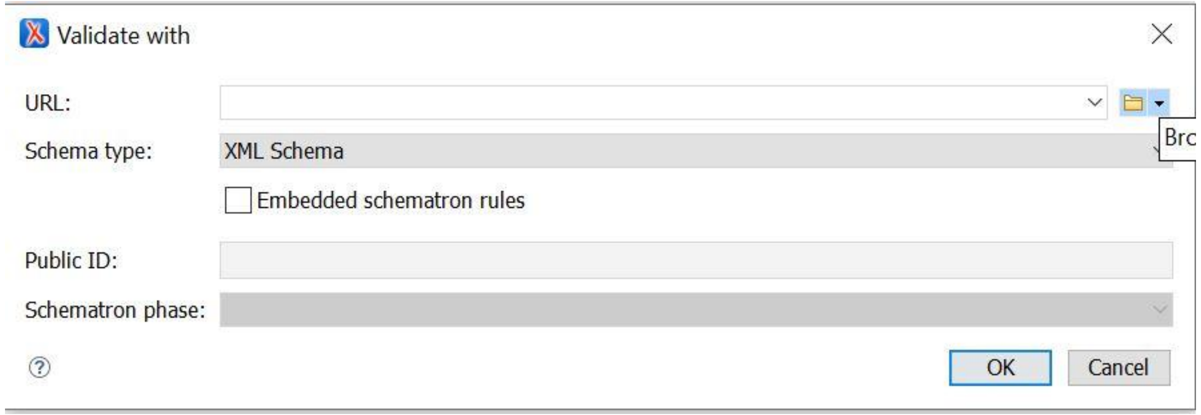
- https://www.oxygenxml.com/
- Paid software vs. 30 days free trial version
    - Academic license: 2 years version control; update annually
- Editing & Developing; Publishing; Collaboration & Review
- Oxygen XML Editor
    - Editing XML files: TEI, MARC21 etc.
    - Editing non-XML files: JSON, HTML, xHTML, CSS, Python, Perl etc.



HTML
Text
Java Programming Language
JavaScript
JSON
JSON Schema
Properties
Perl
C Programming Language
C++ Programming Language
Batch
Shell
PHP
SQL
Python Programming Language

- **Features**
  - XML Editing
    - Check Well-Formedness
    - Validate with DTD, Relax NG, Schematron, different XML schemas (support for XML catalogues → web address or disk path)
    - For documents that do not have a schema, Oxygen can analyze the structure of the document and generate a schema


Validate Ctrl+Shift+V
Validate (cached)
Validate with...
Check Well-Formedness Ctrl+Shift+W
Configure Validation Scenario(s)...
XSV
LIBXML
MSXML4.0
MSXML.NET
Saxon-EE
Configure Custom Validation Engines...


Validate with

URL:

Schema type: XML Schema

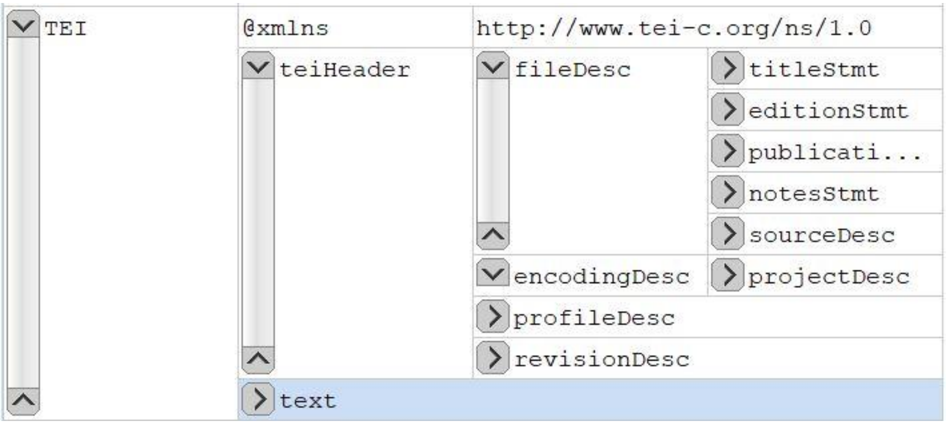
□ Embedded schematron rules

Public ID:

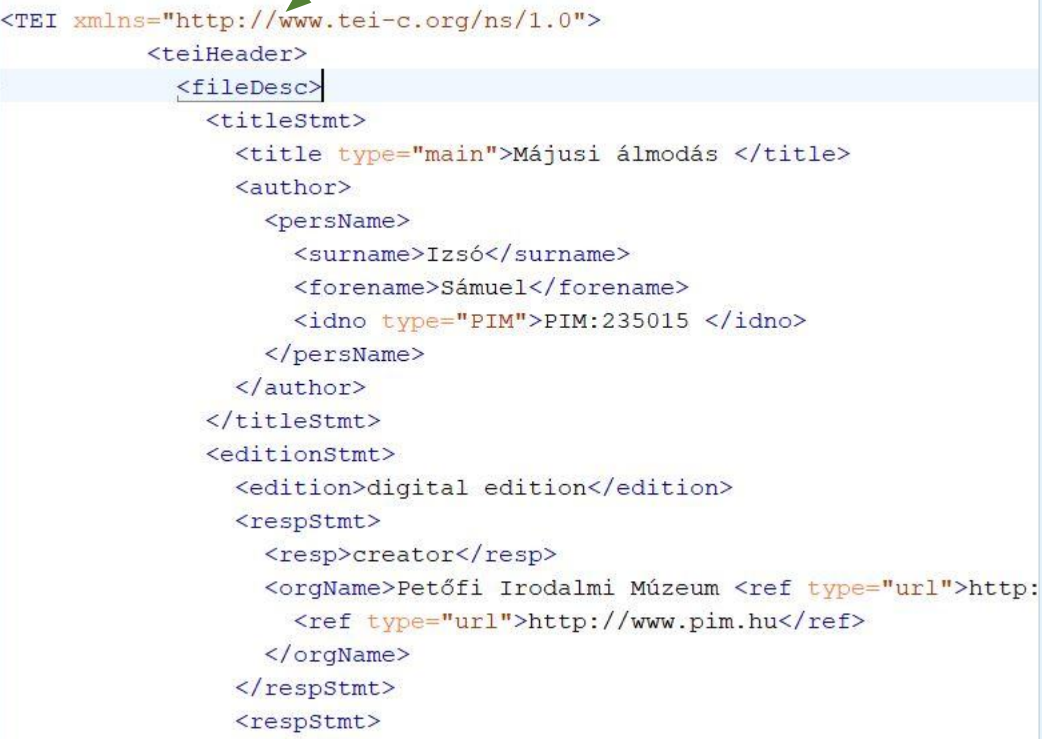
Schematron phase:

Different views

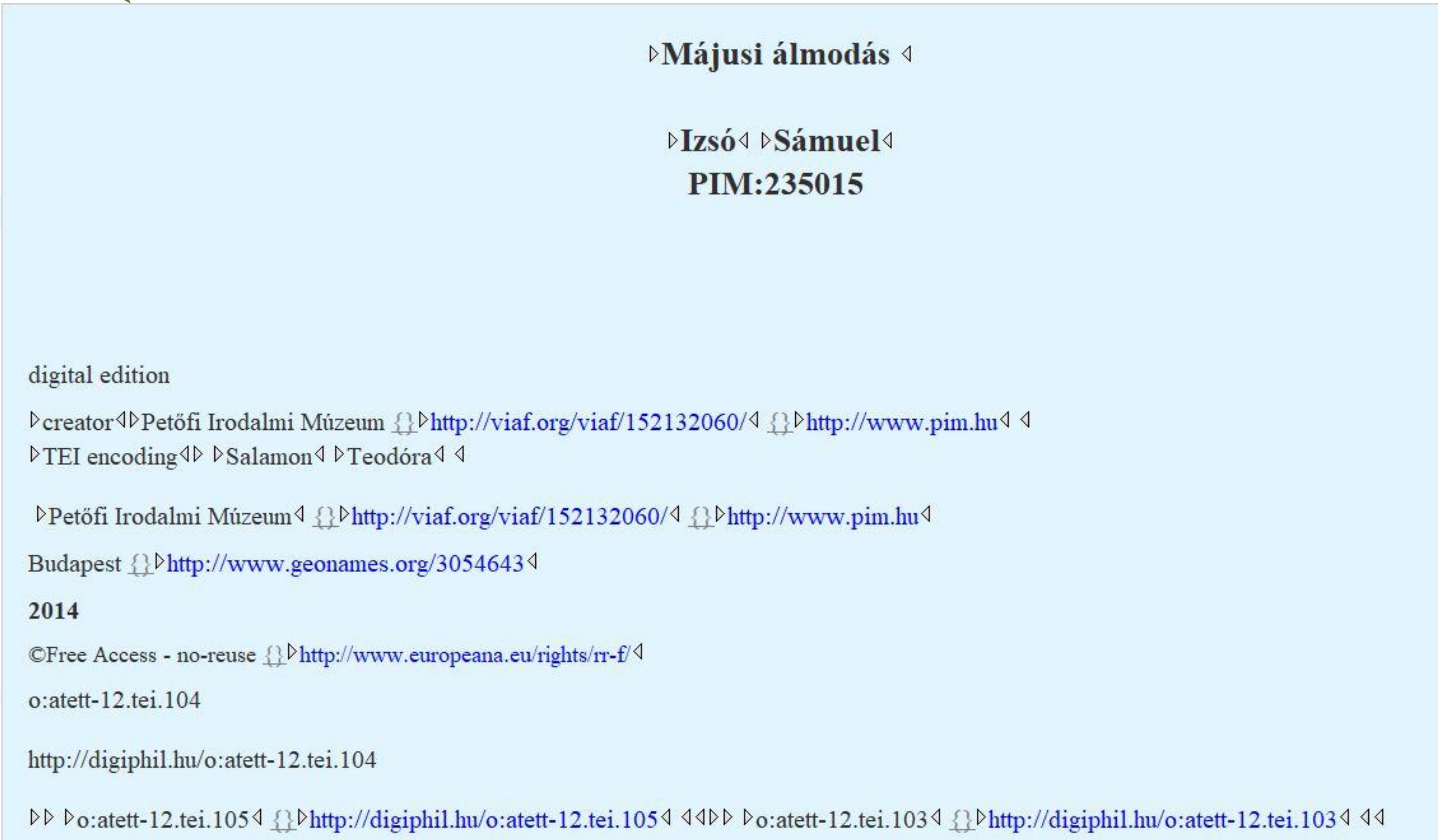
- Grid view: shows the XML document in a spreadsheet-like fashion
- Author view: WYSIWYM (What You See Is What You Mean). This view is based on providing a CSS file for the document. (eg. DITA, DocBook, and TEI)
- Text view: is the default view for editing an XML document.



```
TEI   @xmlns        http://www.tei-c.org/ns/1.0
      teiHeader     fileDesc       titleStmt
                                   editionStmt
                                   publicati...
                                   notesStmt
                                   sourceDesc
                    encodingDesc   projectDesc
                    profileDesc
                    revisionDesc
      text
```



```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
          <teiHeader>
            <fileDesc>
              <titleStmt>
                <title type="main">Májusi álmodás </title>
                <author>
                  <persName>
                    <surname>Izsó</surname>
                    <forename>Sámuel</forename>
                    <idno type="PIM">PIM:235015 </idno>
                  </persName>
                </author>
              </titleStmt>
              <editionStmt>
                <edition>digital edition</edition>
                <respStmt>
                  <resp>creator</resp>
                  <orgName>Petőfi Irodalmi Múzeum <ref type="url">http:
                    <ref type="url">http://www.pim.hu</ref>
                  </orgName>
                </respStmt>
                <respStmt>
```



▷Májusi álmodás ◁

▷Izsó◁ ▷Sámuel◁
PIM:235015

digital edition

▷creator◁▷Petőfi Irodalmi Múzeum {}▷http://viaf.org/viaf/152132060/◁ {}▷http://www.pim.hu◁ ◁
▷TEI encoding◁▷ ▷Salamon◁ ▷Teodóra◁ ◁

▷Petőfi Irodalmi Múzeum◁ {}▷http://viaf.org/viaf/152132060/◁ {}▷http://www.pim.hu◁

Budapest {}▷http://www.geonames.org/3054643◁

2014

©Free Access - no-reuse {}▷http://www.europeana.eu/rights/rr-f/◁

o:atett-12.tei.104

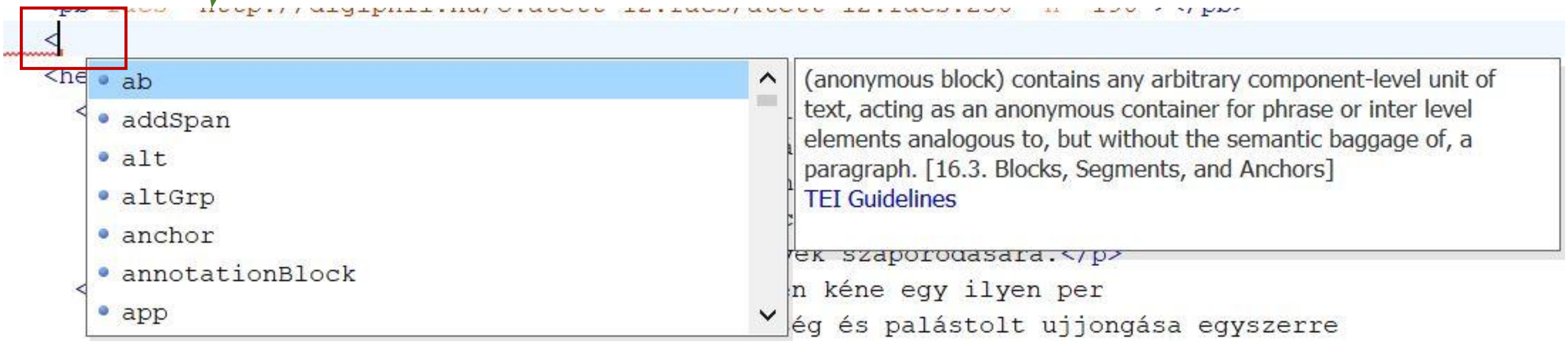
http://digiphil.hu/o:atett-12.tei.104

▷▷ ▷o:atett-12.tei.105◁ {}▷http://digiphil.hu/o:atett-12.tei.105◁ ◁◁▷▷ ▷o:atett-12.tei.103◁ {}▷http://digiphil.hu/o:atett-12.tei.103◁ ◁◁

Text View

Tag completion

Communicate with TEI Guidlines



<

<he

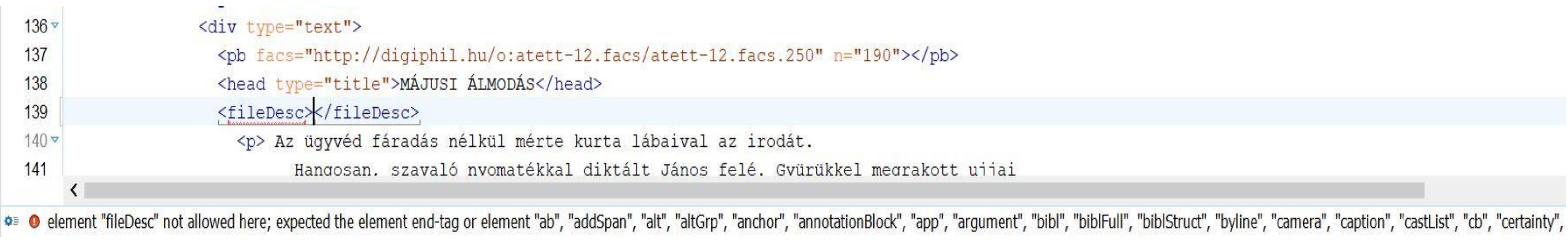
- ab
- addSpan
- alt
- altGrp
- anchor
- annotationBlock
- app

(anonymous block) contains any arbitrary component-level unit of text, acting as an anonymous container for phrase or inter level elements analogous to, but without the semantic baggage of, a paragraph. [16.3. Blocks, Segments, and Anchors]
TEI Guidelines

```
ek szaporodására.</p>
n kéne egy ilyen per
ég és palástolt ujjongása egyszerre
```

Text View – TEI

Validation – Error message



```
<div type="text">
  <pb facs="http://digiphil.hu/o:atett-12.facs/atett-12.facs.250" n="190"></pb>
  <head type="title">MÁJUSI ÁLMODÁS</head>
  <fileDesc></fileDesc>
    <p> Az ügyvéd fáradás nélkül mérte kurta lábaival az irodát.
          Hangosan, szavaló nyomatékkal diktált János felé. Gyürükkel megrakott ujjai
```

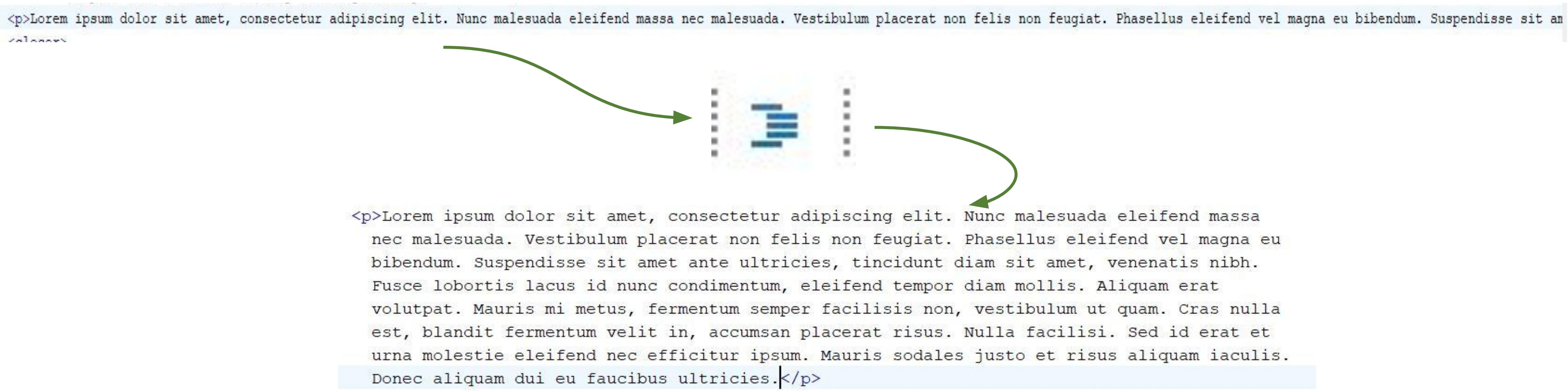
element "fileDesc" not allowed here; expected the element end-tag or element "ab", "addSpan", "alt", "altGrp", "anchor", "annotationBlock", "app", "argument", "bibl", "biblFull", "biblStruct", "byline", "camera", "caption", "castList", "cb", "certainty",

Text View – TEI

Easey to read layout

- Other option: Smart Paste function in Author View

https://www.lipsum.com/feed/html



```
<p>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc malesuada eleifend massa nec malesuada. Vestibulum placerat non felis non feugiat. Phasellus eleifend vel magna eu bibendum. Suspendisse sit an
```

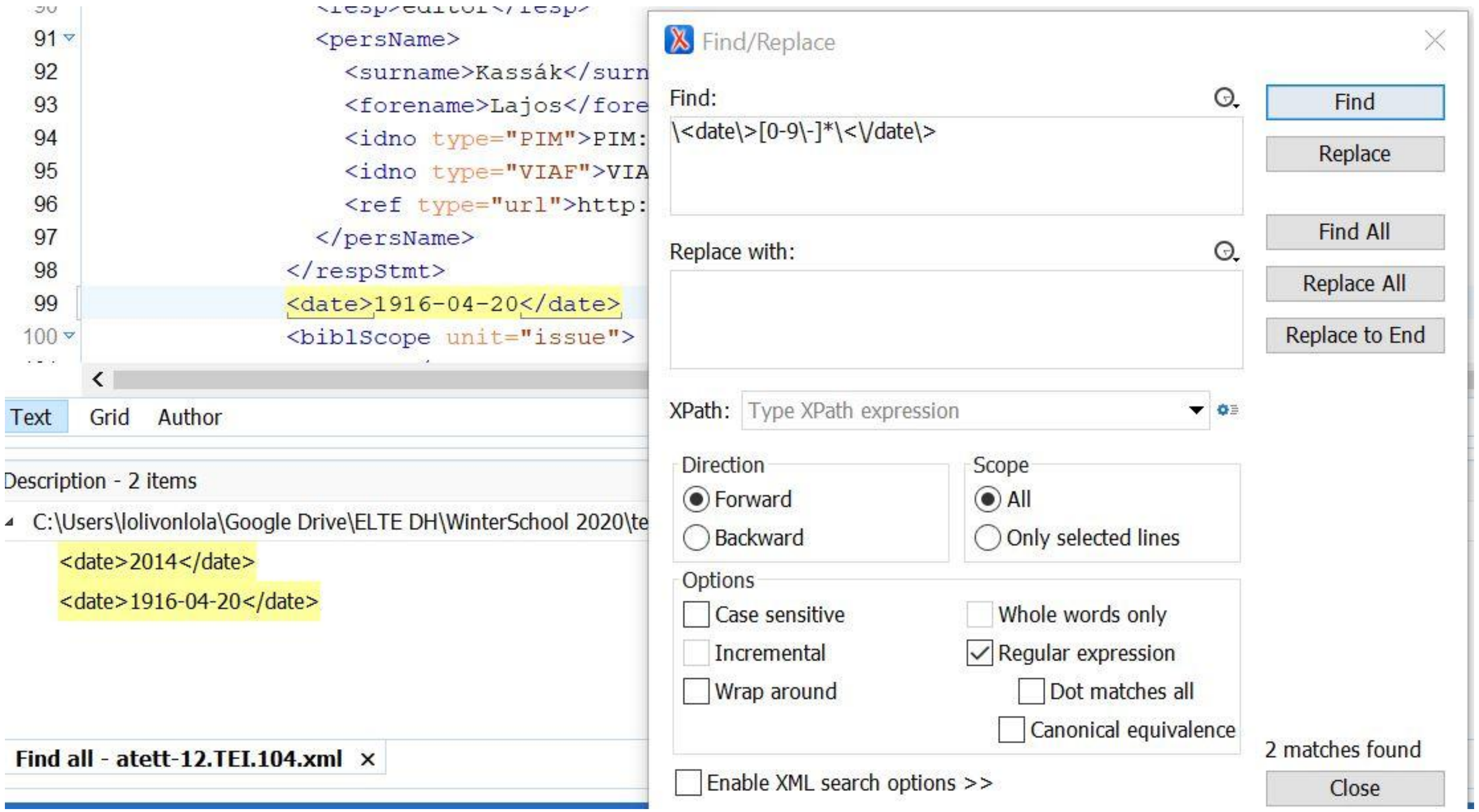
```
<p>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc malesuada eleifend massa
  nec malesuada. Vestibulum placerat non felis non feugiat. Phasellus eleifend vel magna eu
  bibendum. Suspendisse sit amet ante ultricies, tincidunt diam sit amet, venenatis nibh.
  Fusce lobortis lacus id nunc condimentum, eleifend tempor diam mollis. Aliquam erat
  volutpat. Mauris mi metus, fermentum semper facilisis non, vestibulum ut quam. Cras nulla
  est, blandit fermentum velit in, accumsan placerat risus. Nulla facilisi. Sed id erat et
  urna molestie eleifend nec efficitur ipsum. Mauris sodales justo et risus aliquam iaculis.
  Donec aliquam dui eu faucibus ultricies.</p>
```

Features

- Find and Replace
  - Fulltext search
  - Regular expressions

  A regular expression is a sequence of characters that define a search pattern.

    - https://en.wikipedia.org/wiki/Regular_expression
    - Oxygen RegEx syntax: https://www.oxygenxml.com/doc/versions/21.1/ug-editor/topics/regular-expressions.html
    - Opened file (Ctrl + F)
    - Files in a folder (Ctrl + Shift + H)



```
            <resp>editor</resp>
            <persName>
              <surname>Kassák</surn
              <forename>Lajos</fore
              <idno type="PIM">PIM:
              <idno type="VIAF">VIA
              <ref type="url">http:
            </persName>
          </respStmt>
          <date>1916-04-20</date>
          <biblScope unit="issue">
```

Description - 2 items

C:\Users\lolivonlola\Google Drive\ELTE DH\WinterSchool 2020\te

<date>2014</date>

<date>1916-04-20</date>

Find all - atett-12.TEI.104.xml

Find/Replace

Find: `\<date\>[0-9\-]*\<\/date\>`

Replace with:

XPath: Type XPath expression

Direction: ● Forward ○ Backward

Scope: ● All ○ Only selected lines

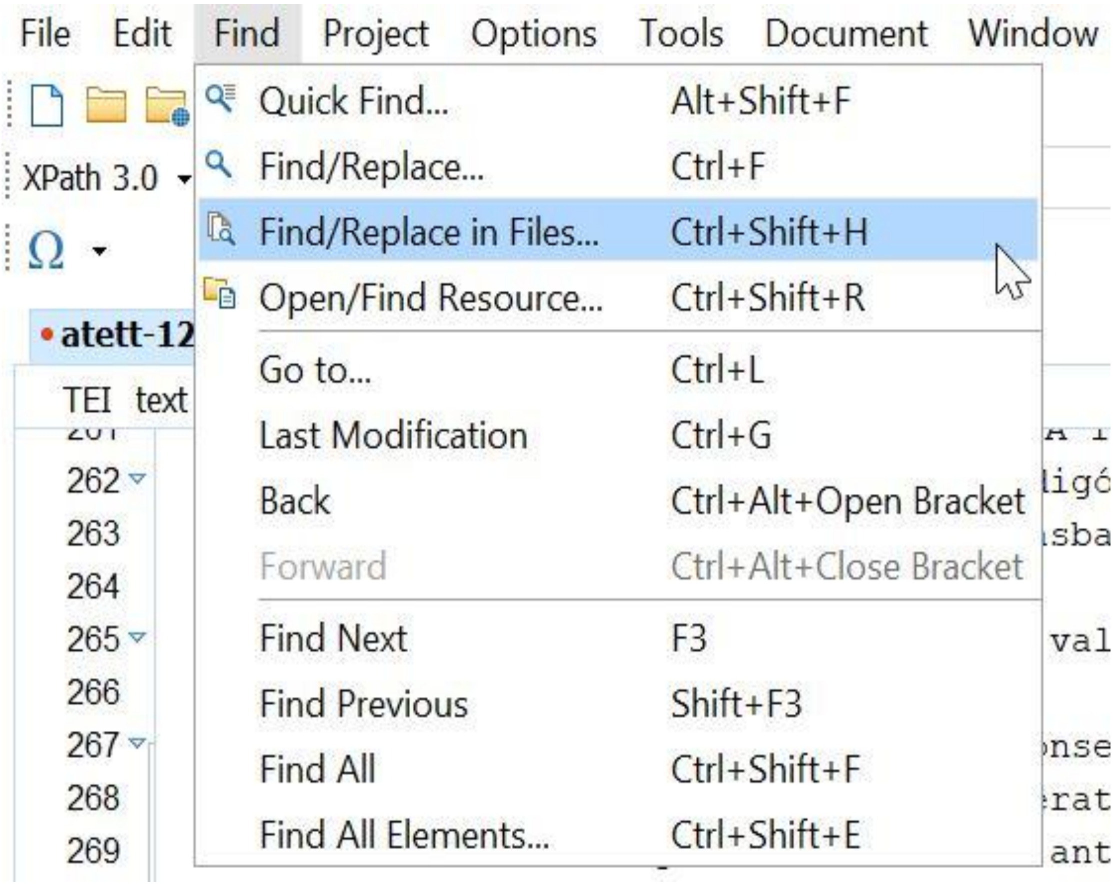
Options:
□ Case sensitive
□ Incremental
□ Wrap around
□ Whole words only
☑ Regular expression
□ Dot matches all
□ Canonical equivalence

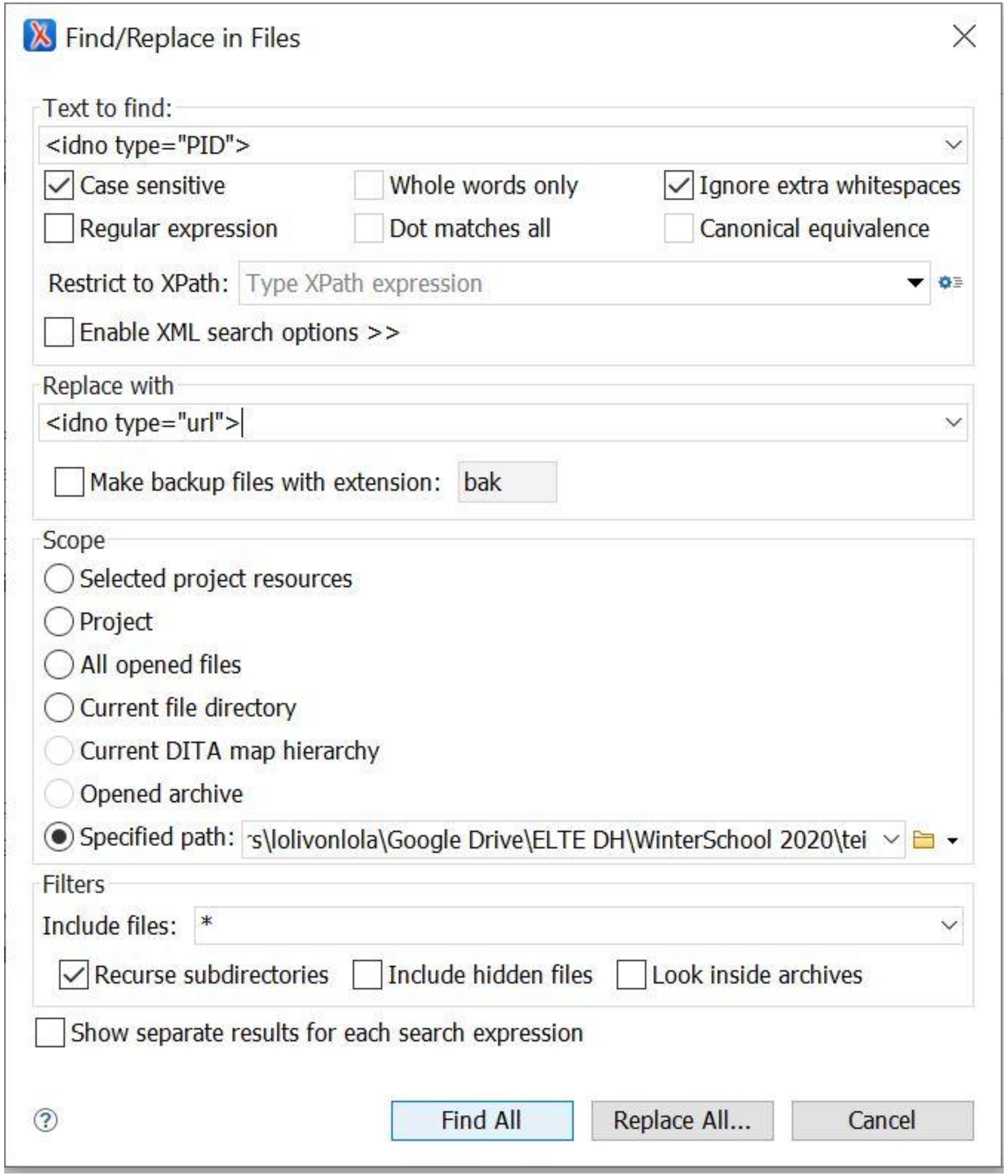
□ Enable XML search options >>

2 matches found

Features

- Files in a folder (Ctrl + Shift + H)


File Edit Find Project Options Tools Document Window
XPath 3.0
Ω
atett-12
TEI text
Quick Find... Alt+Shift+F
Find/Replace... Ctrl+F
Find/Replace in Files... Ctrl+Shift+H
Open/Find Resource... Ctrl+Shift+R
Go to... Ctrl+L
Last Modification Ctrl+G
Back Ctrl+Alt+Open Bracket
Forward Ctrl+Alt+Close Bracket
Find Next F3
Find Previous Shift+F3
Find All Ctrl+Shift+F
Find All Elements... Ctrl+Shift+E



Text to find:
<idno type="PID">

☑ Case sensitive ☐ Whole words only ☑ Ignore extra whitespaces
☐ Regular expression ☐ Dot matches all ☐ Canonical equivalence

Restrict to XPath: Type XPath expression

☐ Enable XML search options >>

Replace with
<idno type="url">

☐ Make backup files with extension: bak

Scope
○ Selected project resources
○ Project
○ All opened files
○ Current file directory
○ Current DITA map hierarchy
○ Opened archive
● Specified path: s\lolivonlola\Google Drive\ELTE DH\WinterSchool 2020\tei

Filters
Include files: *
☑ Recurse subdirectories ☐ Include hidden files ☐ Look inside archives

☐ Show separate results for each search expression

Oxygen XML Editor - Projects

- helps you organize your XML-related files into projects
- batch operations over sets of files
  - Find and Replace
  - XPath in files
  - XQuery in files
  - XSLT transformation

XPath

- is a query language for selecting nodes from an XML document
- may be used to compute values (e.g., strings, numbers, or Boolean values) from the content of an XML document
- based on a tree representation of the XML document, and provides the ability to navigate around the tree
- https://www.w3schools.com/xml/xpath_syntax.asp

Operators:

- /, // and [...] operators, used in path expressions
- Union operator, |, which forms the union of two node-sets.
- Boolean operators "and" and "or", and a function "not()"
- Arithmetic operators +, -, *, "div" (divide), and "mod"
- Comparison operators =, !=, <, >, <=, >=

XPath

Functions

Functions to **manipulate strings**: concat(), substring(), contains(), substring-before(), substring-after(), translate(), normalize-space(), string-length()

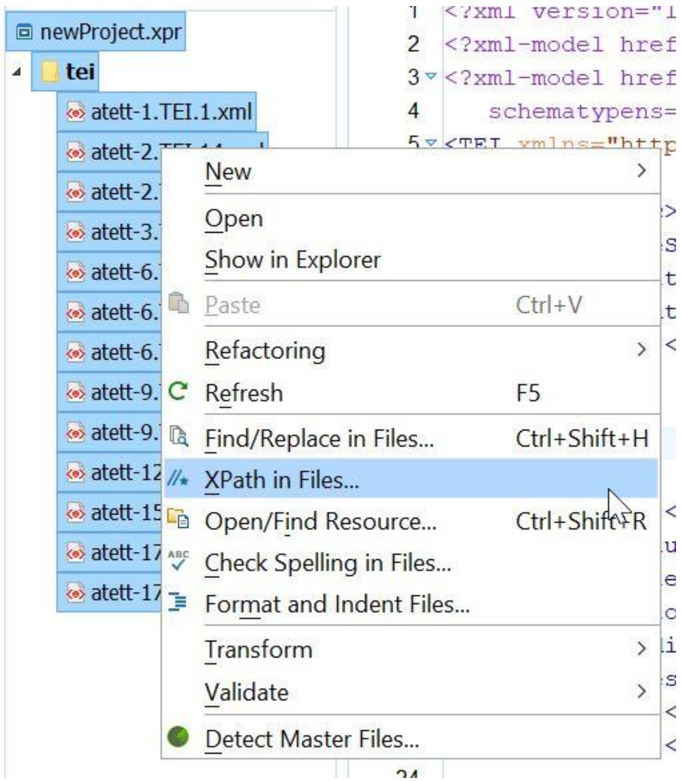
Functions to **manipulate numbers**: sum(), round(), floor(), ceiling()

Functions to **get properties of nodes**: name(), local-name(), namespace-uri()

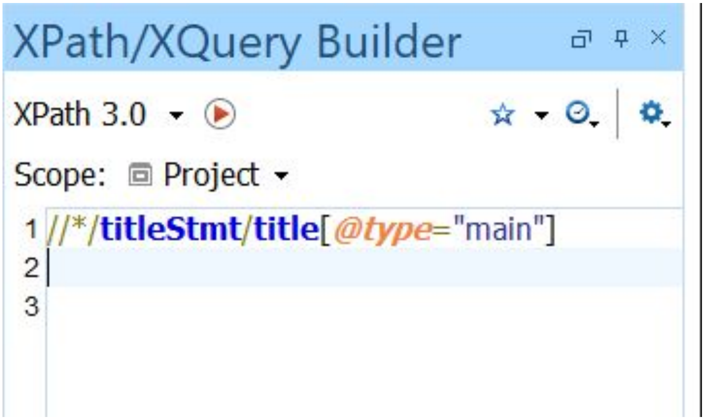
Functions to **get information about the processing context**: position(), last()

Type **conversion functions**: string(), number(), boolean()

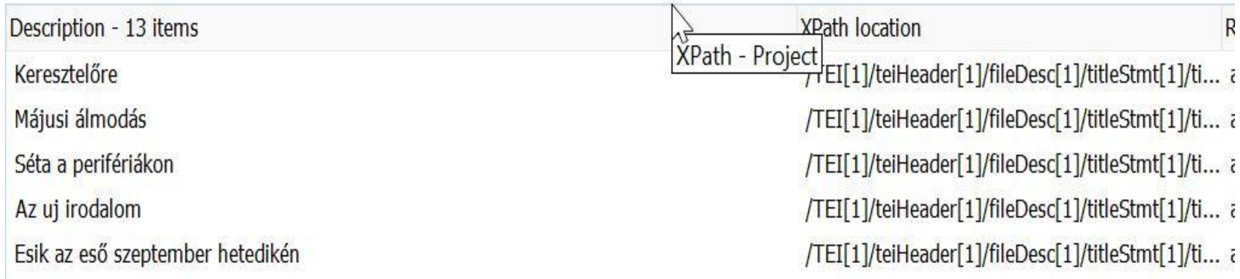
- Boolean operators
  - not(), true(), false()
- Example:



```
<?xml version="1
<?xml-model href
<?xml-model href
   schematypens=
<TEI xmlns="http
```

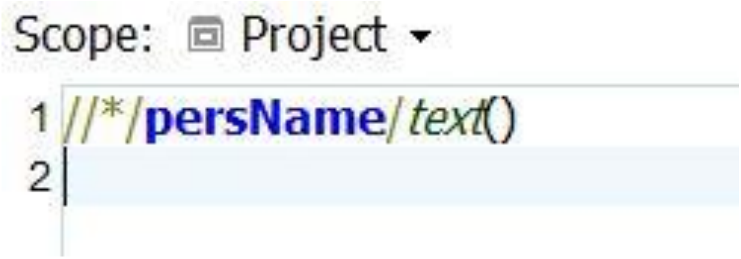


```
//*/titleStmt/title[@type="main"]
```



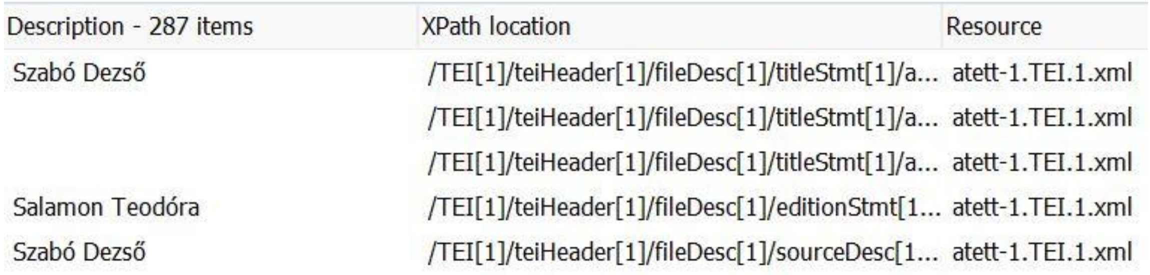
Description - 13 items	XPath location	R
Keresztelőre	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/ti...	a
Májusi álmodás	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/ti...	a
Séta a perifériákon	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/ti...	a
Az uj irodalom	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/ti...	a
Esik az eső szeptember hetedikén	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/ti...	a

- Example:

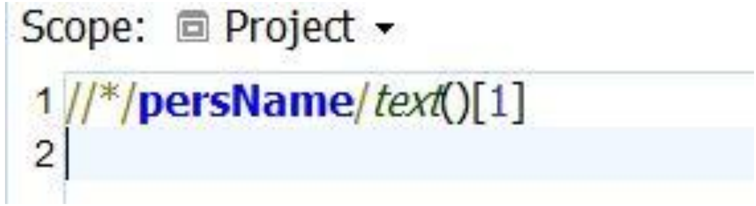


Scope: Project

```
//*/persName/text()
```

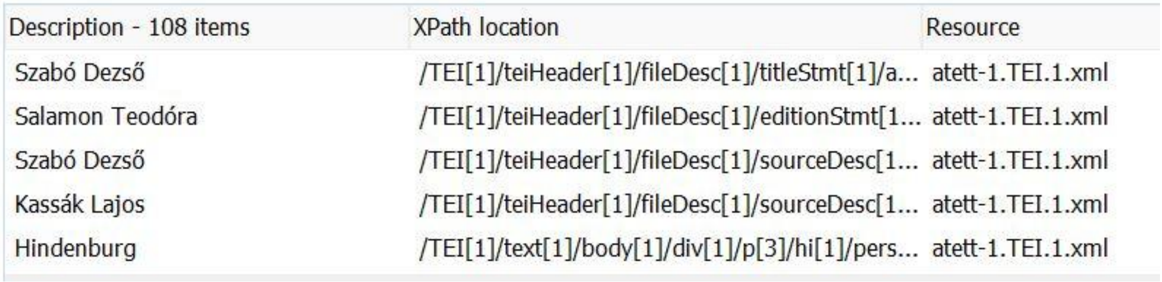


Description - 287 items	XPath location	Resource
Szabó Dezső	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/a...	atett-1.TEI.1.xml
	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/a...	atett-1.TEI.1.xml
	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/a...	atett-1.TEI.1.xml
Salamon Teodóra	/TEI[1]/teiHeader[1]/fileDesc[1]/editionStmt[1...	atett-1.TEI.1.xml
Szabó Dezső	/TEI[1]/teiHeader[1]/fileDesc[1]/sourceDesc[1...	atett-1.TEI.1.xml

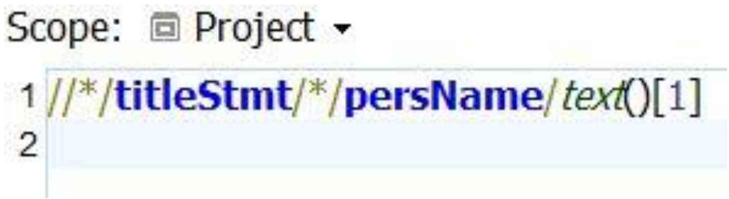


Scope: Project

```
//*/persName/text()[1]
```



Description - 108 items	XPath location	Resource
Szabó Dezső	/TEI[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]/a...	atett-1.TEI.1.xml
Salamon Teodóra	/TEI[1]/teiHeader[1]/fileDesc[1]/editionStmt[1...	atett-1.TEI.1.xml
Szabó Dezső	/TEI[1]/teiHeader[1]/fileDesc[1]/sourceDesc[1...	atett-1.TEI.1.xml
Kassák Lajos	/TEI[1]/teiHeader[1]/fileDesc[1]/sourceDesc[1...	atett-1.TEI.1.xml
Hindenburg	/TEI[1]/text[1]/body[1]/div[1]/p[3]/hi[1]/pers...	atett-1.TEI.1.xml



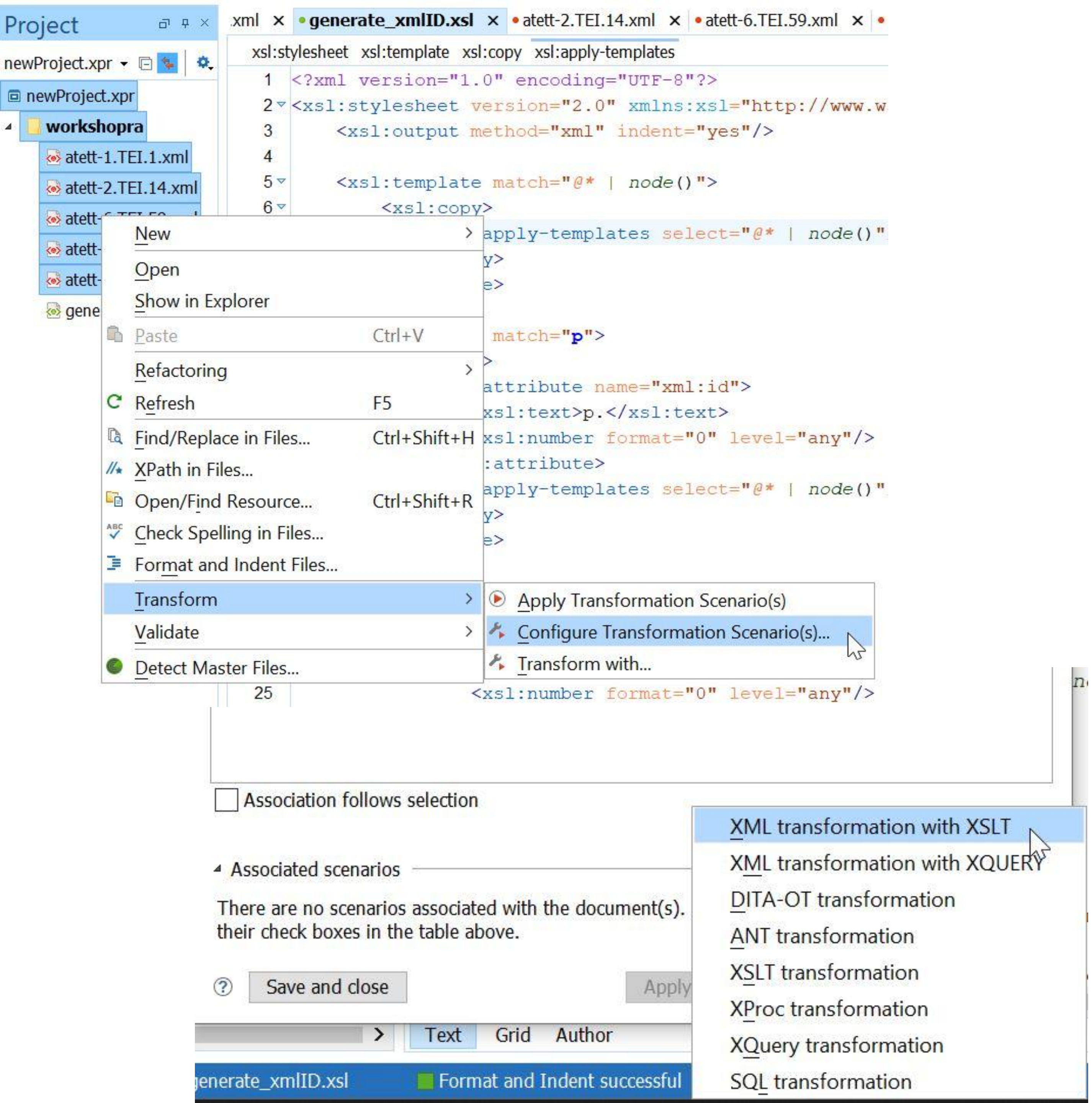
Scope: Project

```
//*/titleStmt/*/persName/text()[1]
```

- Excercise: Please list every value of the <idno> element with the @type „PID”.
- Excersice: Please list every value of the <idno> element with the @type „PID” AND the value of the @prev attribute.

Extensible Stylesheet Language Transformations

- language for transforming XML documents into other XML documents, or other formats such as HTML, plain text, PDF etc.
- new file is created, the original file is not changed
- XSLT uses XPath to identify subsets of the source document tree and perform calculations
- Example:
  - generate @xml:id automatically



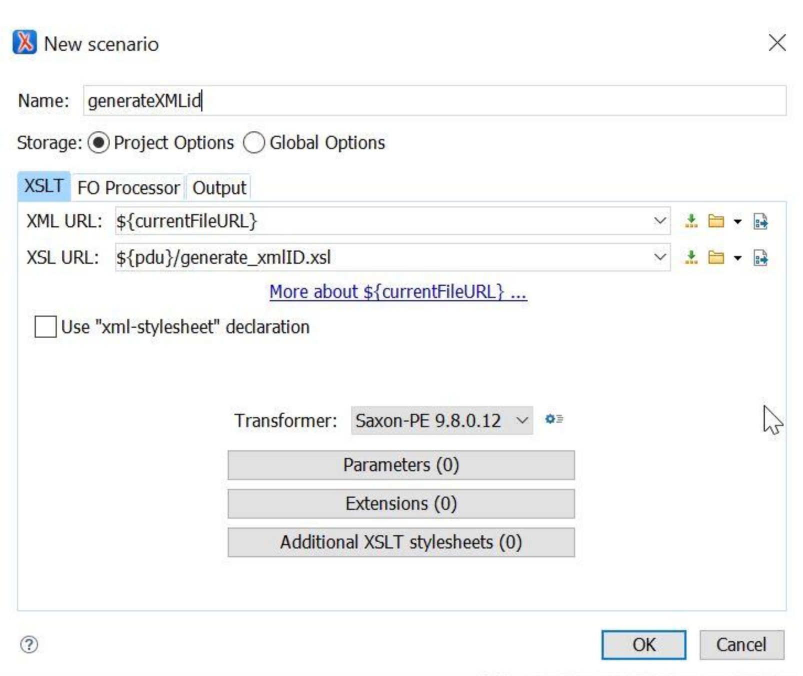
```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="2.0" xmlns:xsl="http://www.w
    <xsl:output method="xml" indent="yes"/>

    <xsl:template match="@* | node()">
        <xsl:copy>
apply-templates select="@* | node()"
y>
e>

 match="p">
>
attribute name="xml:id">
xsl:text>p.</xsl:text>
xsl:number format="0" level="any"/>
:attribute>
apply-templates select="@* | node()"
y>
e>
            <xsl:number format="0" level="any"/>
```

☐ Association follows selection

Associated scenarios

There are no scenarios associated with the document(s).
their check boxes in the table above.

enerate_xmlID.xsl   Format and Indent successful



Name: generateXMLid

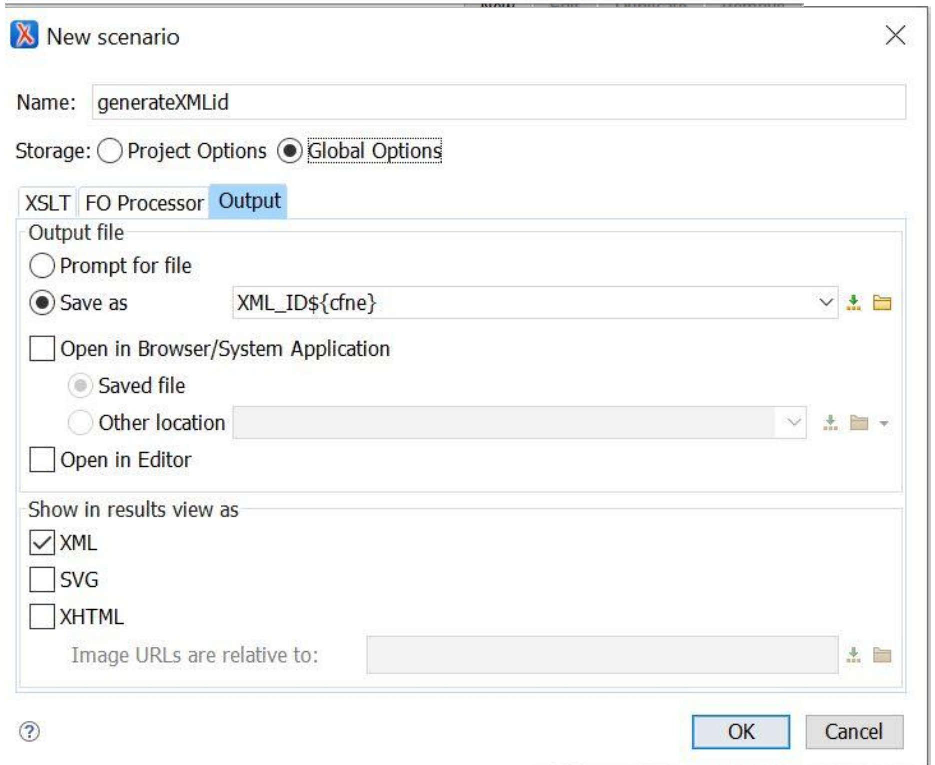
Storage: ● Project Options ○ Global Options

XML URL: ${currentFileURL}

XSL URL: ${pdu}/generate_xmlID.xsl

☐ Use "xml-stylesheet" declaration

Transformer: Saxon-PE 9.8.0.12



Name: generateXMLid

Storage: ○ Project Options ● Global Options

Output file

○ Prompt for file

● Save as XML_ID${cfne}

☐ Open in Browser/System Application

○ Saved file

○ Other location

☐ Open in Editor

Show in results view as

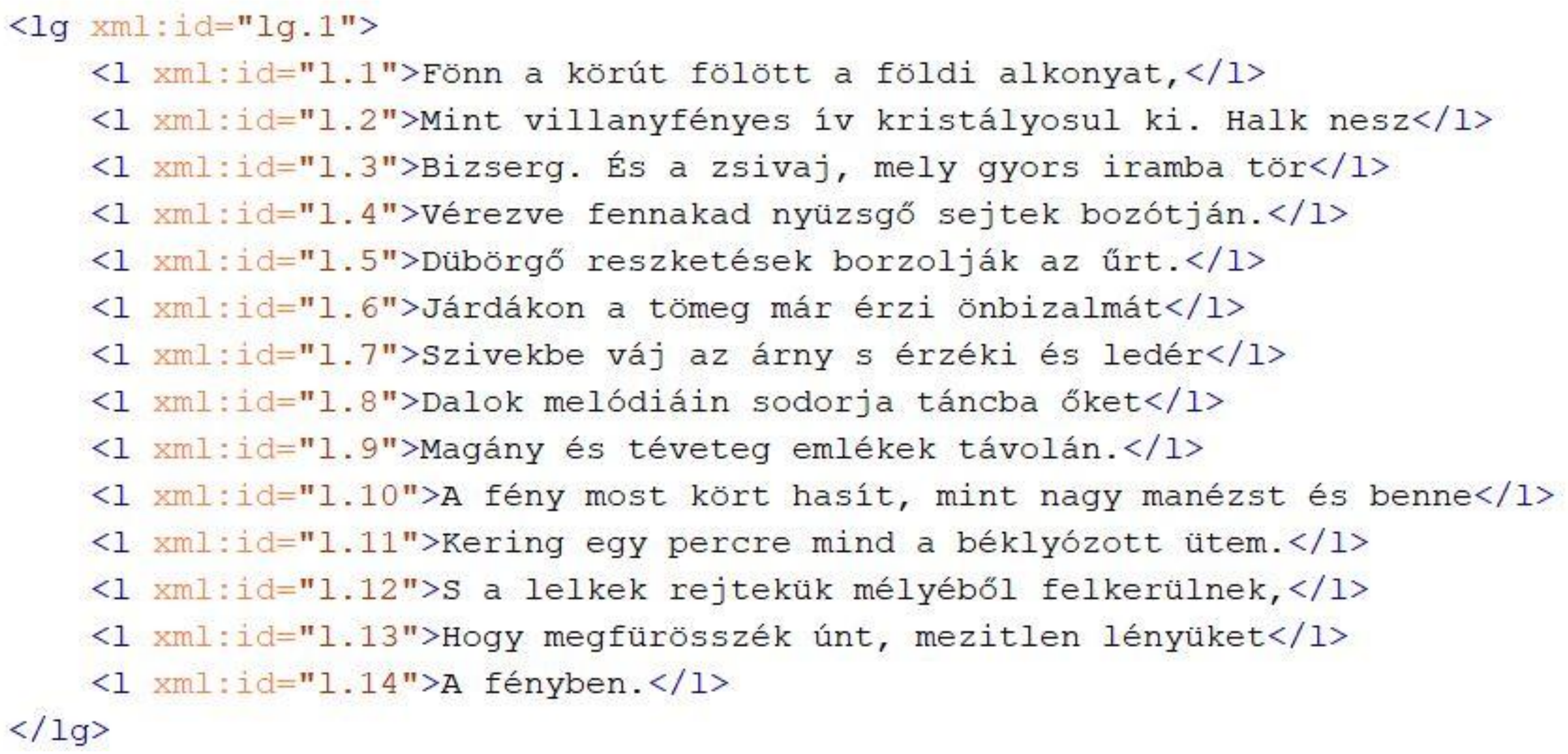
☑ XML

☐ SVG

☐ XHTML

Image URLs are relative to:

Result



```
<lg xml:id="lg.1">
    <l xml:id="l.1">Fönn a körút fölött a földi alkonyat,</l>
    <l xml:id="l.2">Mint villanyfényes ív kristályosul ki. Halk nesz</l>
    <l xml:id="l.3">Bizserg. És a zsivaj, mely gyors iramba tör</l>
    <l xml:id="l.4">Vérezve fennakad nyüzsgő sejtek bozótján.</l>
    <l xml:id="l.5">Dübörgő reszketések borzolják az űrt.</l>
    <l xml:id="l.6">Járdákon a tömeg már érzi önbizalmát</l>
    <l xml:id="l.7">Szivekbe váj az árny s érzéki és ledér</l>
    <l xml:id="l.8">Dalok melódiáin sodorja táncba őket</l>
    <l xml:id="l.9">Magány és téveteg emlékek távolán.</l>
    <l xml:id="l.10">A fény most kört hasít, mint nagy manézst és benne</l>
    <l xml:id="l.11">Kering egy percre mind a béklyózott ütem.</l>
    <l xml:id="l.12">S a lelkek rejtekük mélyéből felkerülnek,</l>
    <l xml:id="l.13">Hogy megfürösszék únt, mezitlen lényüket</l>
    <l xml:id="l.14">A fényben.</l>
</lg>
```