Modelling humanities data with TEI-XML

SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Dr Katarzyna Anna Kapitan
19 November 2025

XPath

XML PATH LANGUAGE

XPath

- XPath (XML Path Language) is W3C (World Wide Web Consortium) recommendation
- XPath uses "path like" syntax to identify and navigate nodes in XML docs
- XPath is used for navigating through elements and attributes in XML docs
  - This means we use it for finding things in XML documents when, for example, we prepare data exports and/or transformations.

XPath in XSLT

- XPath is a major element in the XSLT standard.
- XPath contains **over 200** built-in functions, which beyond finding elements and nodes also allow you to perform various operations (as counting items and summing values).
- Below one section from the XSLT file from last week that we used to extract relevant info from Dares, chapter 14.

```
 name="totalShipNumber" select="sum(text//num/@value)"/>
 name="totalPlaces" select="count(text//placeName)"/>
from <xsl:value-of select="$totalPlaces"/> cities in Greece a
xsl:value-of select="$totalShipNumber"/> ships.</p>
```

XPath Syntax: Expressions

/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

https://www.w3schools.com/xml/xpath_syntax.asp

/bookstore	Selects the root element bookstore **Note:** If the path starts with a slash (/) it always represents an absolute path to an element!
bookstore/book	Selects all book elements that are children of bookstore
//book	Selects all book elements no matter where they are in the document
bookstore//book	Selects all book elements that are descendant of the bookstore element, no matter where they are under the bookstore element
//@lang	Selects all attributes that are named lang

https://www.w3schools.com/xml/xpath_syntax.asp

XPath Syntax: Wildcards

Wildcard	Description
*	Matches any element node
@*	Matches any attribute node
node()	Matches any node of any kind

https://www.w3schools.com/xml/xpath_syntax.asp

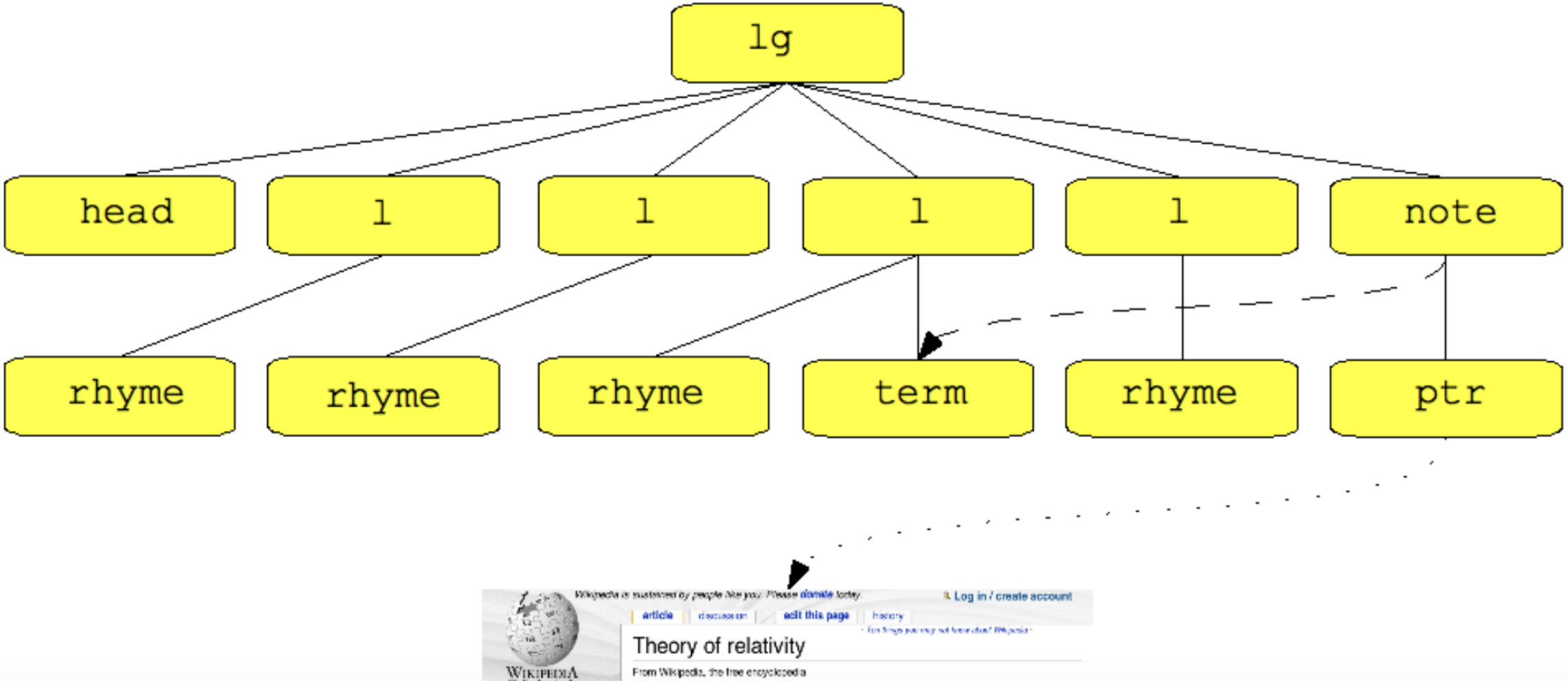
Sample document instance

```
<?xml version="1.0" encoding="UTF-8"?>
<lg type="limerick" rhyme="aabba" n="3">
  <head>Warp Speed, Ms Bright!</head>
  <l>There was a young lady named <rhyme label="a">Bright</rhyme>,</l>
  <l>Who travelled much faster than <rhyme label="a">light</rhyme>,</l>
  <l>She departed one <rhyme label="b">day</rhyme>,</l>
  <l>In a <term xml:id="t17">relative</term> way <rhyme label="b">way</rhyme>,</l>
  <l>And returned on the previous <rhyme label="a">night</rhyme>.</l>
  <note target="#t17">See
    <ptr target="http://en.wikipedia.org/wiki/Theory_of_relativity"/>.</note>
</lg>
```

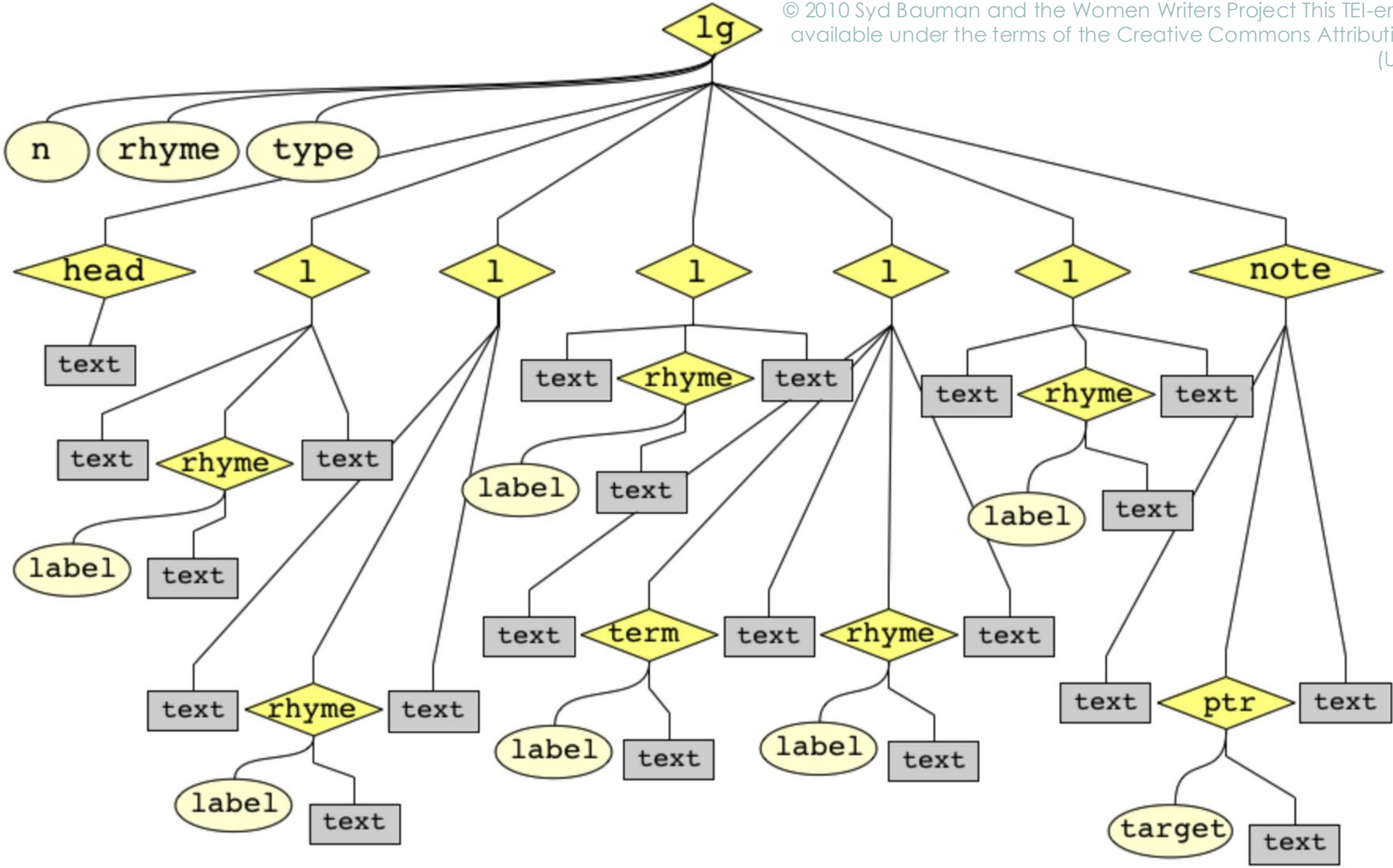
© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

Simplified XML tree

© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.



© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

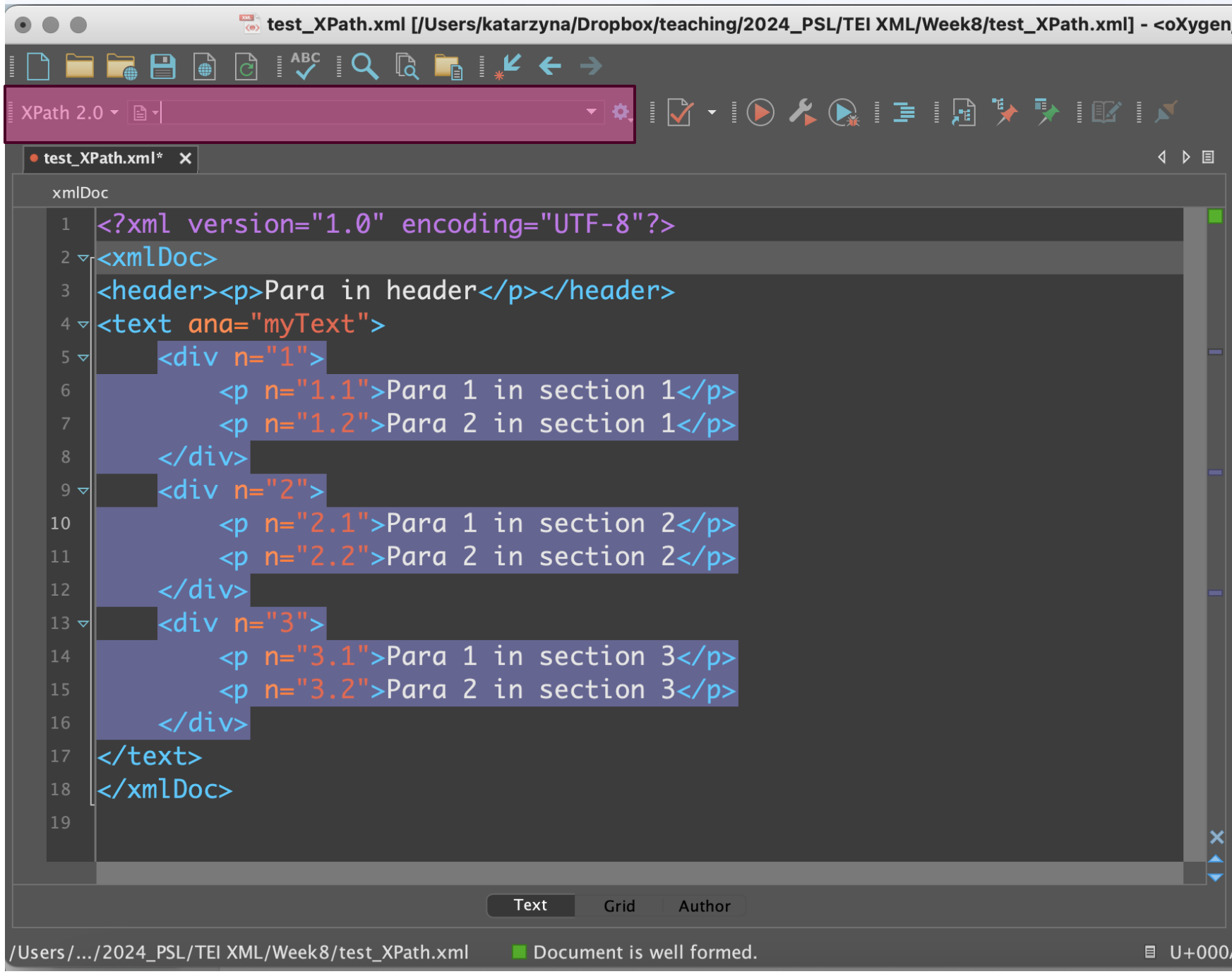


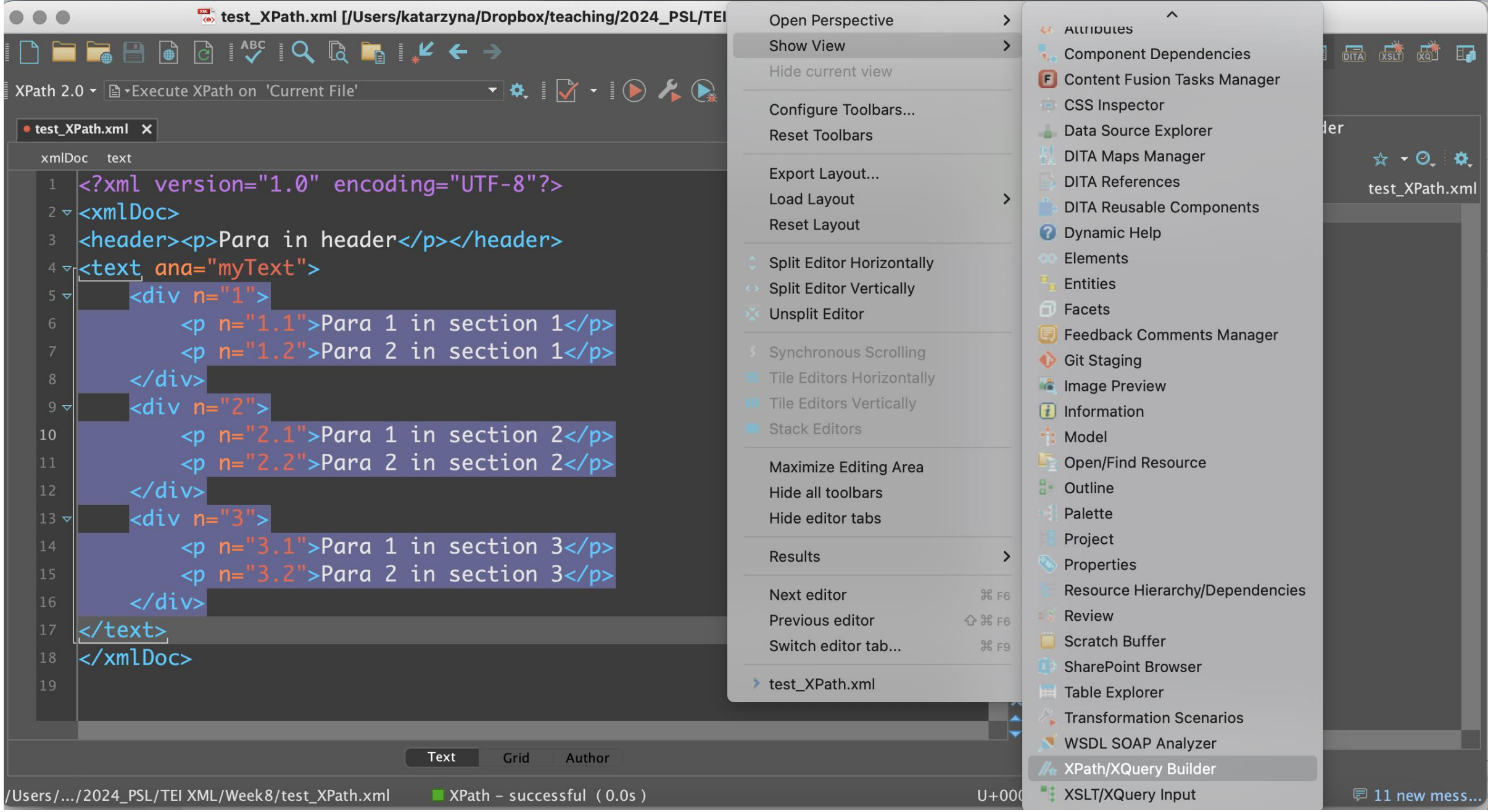
Exercise 1: Testing XPath in Oxygen

Open the following file with Oxygen XML Editor:
test_Xpath.xml

GitHub: Week8/Exercises/Ex1

XPath 2.0


test_XPath.xml [/Users/katarzyna/Dropbox/teaching/2024_PSL/TEI XML/Week8/test_XPath.xml] - <oXygen
XPath 2.0
test_XPath.xml*
xmlDoc
1 <?xml version="1.0" encoding="UTF-8"?>
2 <xmlDoc>
3 <header> Para in header </header>
4 <text ana="myText">
5
6 Para 1 in section 1
7 Para 2 in section 1
8
9
10 Para 1 in section 2
11 Para 2 in section 2
12
13
14 Para 1 in section 3
15 Para 2 in section 3
16
17 </text>
18 </xmlDoc>
19
Text
Grid
Author
/Users/.../2024_PSL/TEI XML/Week8/test_XPath.xml
Document is well formed.


test_XPath.xml [/Users/katarzyna/Dropbox/teaching/2024_PSL/TEI
XPath 2.0
Execute XPath on 'Current File'
test_XPath.xml
xmlDoc
text
<?xml version="1.0" encoding="UTF-8"?>
<xmlDoc>
<header> Para in header </header>
<text ana="myText">

Para 1 in section 1
Para 2 in section 1

Para 1 in section 2
Para 2 in section 2

Para 1 in section 3
Para 2 in section 3

</text>
</xmlDoc>
Text
Grid
Author
/Users/.../2024_PSL/TEI XML/Week8/test_XPath.xml
XPath - successful (0.0s)
Open Perspective
Show View
Hide current view
Configure Toolbars...
Reset Toolbars
Export Layout...
Load Layout
Reset Layout
Split Editor Horizontally
Split Editor Vertically
Unsplit Editor
Synchronous Scrolling
Tile Editors Horizontally
Tile Editors Vertically
Stack Editors
Maximize Editing Area
Hide all toolbars
Hide editor tabs
Results
Next editor
Previous editor
Switch editor tab...
test_XPath.xml
Component Dependencies
Content Fusion Tasks Manager
CSS Inspector
Data Source Explorer
DITA Maps Manager
DITA References
DITA Reusable Components
Dynamic Help
Elements
Entities
Facets
Feedback Comments Manager
Git Staging
Image Preview
Information
Model
Open/Find Resource
Outline
Palette
Project
Properties
Resource Hierarchy/Dependencies
Review
Scratch Buffer
SharePoint Browser
Table Explorer
Transformation Scenarios
WSDL SOAP Analyzer
XPath/XQuery Builder
XSLT/XQuery Input
11 new mess...

Xpath: //p

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xmlDoc>
<header><p>Para in header</p></header>
<text ana="myText">
    <div n="1">
        <p n="1.1">Para 1 in section 1</p>
        <p n="1.2">Para 2 in section 1</p>
    </div>
    <div n="2">
        <p n="2.1">Para 1 in section 2</p>
        <p n="2.2">Para 2 in section 2</p>
    </div>
    <div n="3">
        <p n="3.1">Para 1 in section 3</p>
        <p n="3.2">Para 2 in section 3</p>
    </div>
</text>
</xmlDoc>
```

//div/p

```
<?xml version="1.0" encoding="UTF-8"?>
<xmlDoc>
<header><p>Para in header</p></header>
<text ana="myText">
    <div n="1">
        <p n="1.1">Para 1 in section 1</p>
        <p n="1.2">Para 2 in section 1</p>
    </div>
    <div n="2">
        <p n="2.1">Para 1 in section 2</p>
        <p n="2.2">Para 2 in section 2</p>
    </div>
    <div n="3">
        <p n="3.1">Para 1 in section 3</p>
        <p n="3.2">Para 2 in section 3</p>
    </div>
</text>
</xmlDoc>
```

//p//text()

test_XPath.xml

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <xmlDoc>
3  <header><p>Para in header</p></header>
4  <text ana="myText">
5      <div n="1">
6          <p n="1.1">Para 1 in section 1</p>
7          <p n="1.2">Para 2 in section 1</p>
8      </div>
9      <div n="2">
10         <p n="2.1">Para 1 in section 2</p>
11         <p n="2.2">Para 2 in section 2</p>
12     </div>
13     <div n="3">
14         <p n="3.1">Para 1 in section 3</p>
15         <p n="3.2">Para 2 in section 3</p>
16     </div>
17 </text>
18 </xmlDoc>
19
```

//text/*

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xmlDoc>
<header><p>Para in header</p></header>
<text ana="myText">
    <div n="1">
        <p n="1.1">Para 1 in section 1</p>
        <p n="1.2">Para 2 in section 1</p>
    </div>
    <div n="2">
        <p n="2.1">Para 1 in section 2</p>
        <p n="2.2">Para 2 in section 2</p>
    </div>
    <div n="3">
        <p n="3.1">Para 1 in section 3</p>
        <p n="3.2">Para 2 in section 3</p>
    </div>
</text>
</xmlDoc>
```

Translate to prose
(What does each statement mean?

1. //title
2. //book/title
3. //chapter/footnote
4. //chapter//footnote
5. What's the difference between 3 and 4?

Predicates

- If you only want Act 3, Scene 1:

```
/TEI/text/body/div[3]/div[1]
```

- Works well presuming you know what you want by element count.
- But in many cases, that is at least inconvenient, if not outright unknown.
- No matter how many s there are, we know this scene has the identifier "sha-ham301" Thus:

```
//div[ @xml:id = 'sha-ham301' ]
```

selects the same node.

© 2010 Syd Bauman and the Women Writers Project This TEI-encoded
available under the terms of the Creative Commons Attribution-Shar
(Unporte

Predicates

XPath	selects
`//listPlace/place[1]`	the first <place> of each <listPlace> (of which there only happens to be one)
`//*[@cRef]`	all elements that have a cRef= attribute
`//title[@level='m']`	all monographic titles
`/TEI/text//name[ not( @key ) ]`	<name> elements that are missing their key= attributes
`//lg[@type='song']/l[1]`	list first line of each song (16 nodes)
`(//lg[@type='song']/l)[1]`	returns first line of all songs (1 node)

© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

//p[@n="2.1"]

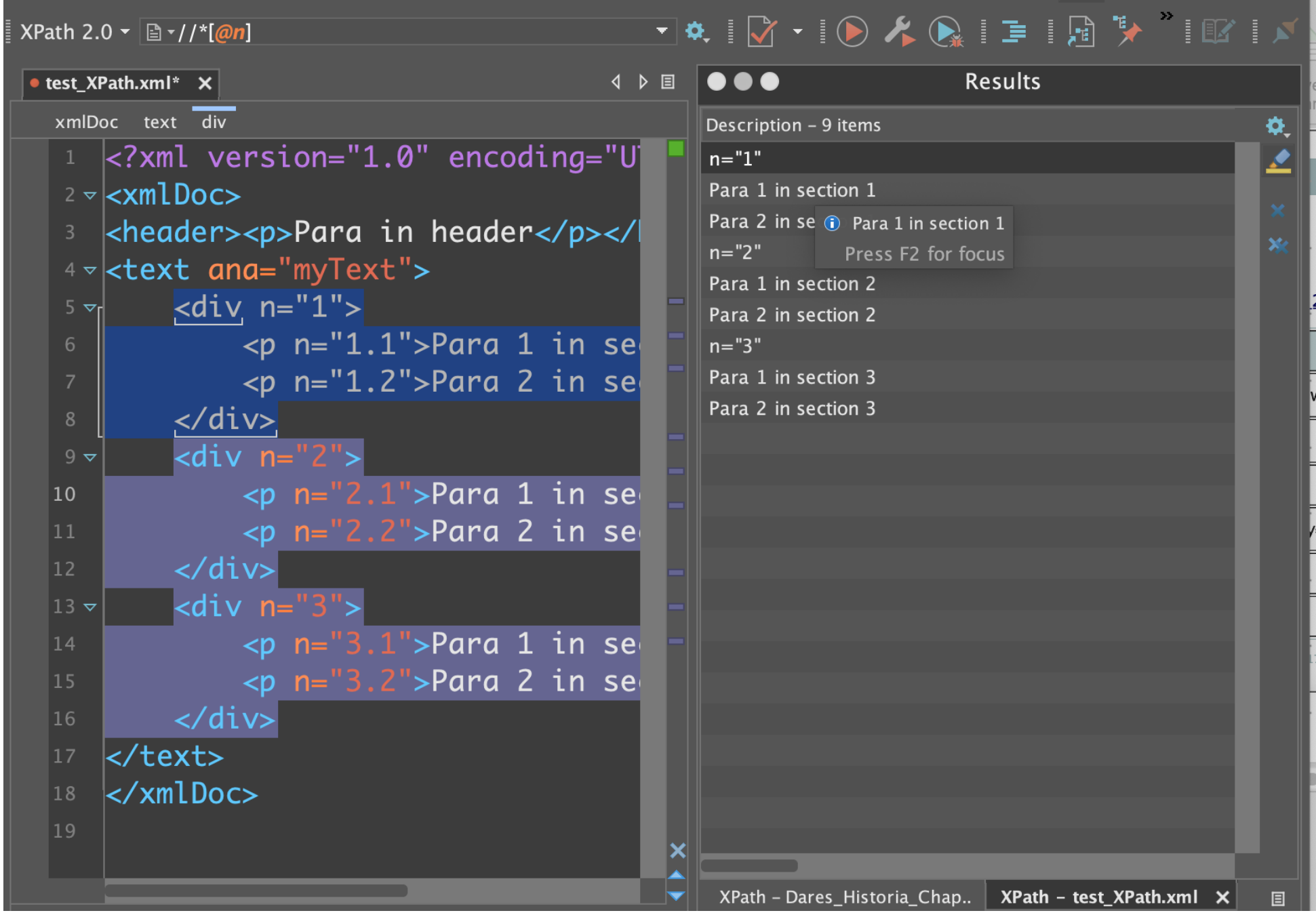
XPath 2.0 //p[@n="2.1"]

test_XPath.xml*

xmlDoc text div p

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <xmlDoc>
3  <header><p>Para in header</p></header>
4  <text ana="myText">
5      <div n="1">
6          <p n="1.1">Para 1 in section 1</p>
7          <p n="1.2">Para 2 in section 1</p>
8      </div>
9      <div n="2">
10         <p n="2.1">Para 1 in section 2</p>
11         <p n="2.2">Para 2 in section 2</p>
12     </div>
13     <div n="3">
14         <p n="3.1">Para 1 in section 3</p>
15         <p n="3.2">Para 2 in section 3</p>
16     </div>
17 </text>
18 </xmlDoc>
19
```

//*[@n]


XPath 2.0
//*[@n]
test_XPath.xml*
xmlDoc
text
div
<?xml version="1.0" encoding="U
<xmlDoc>
<header> Para in header </
<text ana="myText">

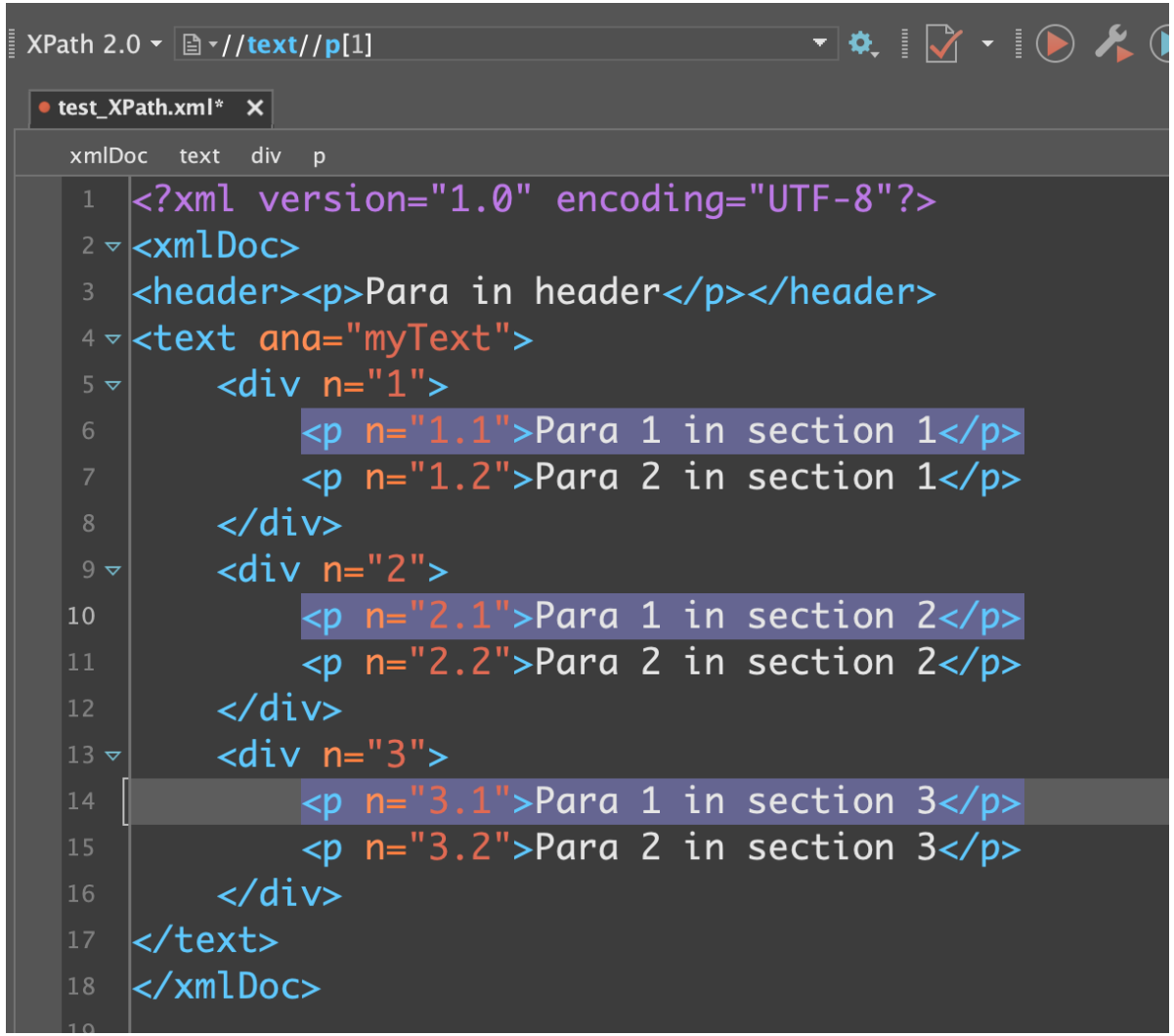
Para 1 in se
Para 2 in se

Para 1 in se
Para 2 in se

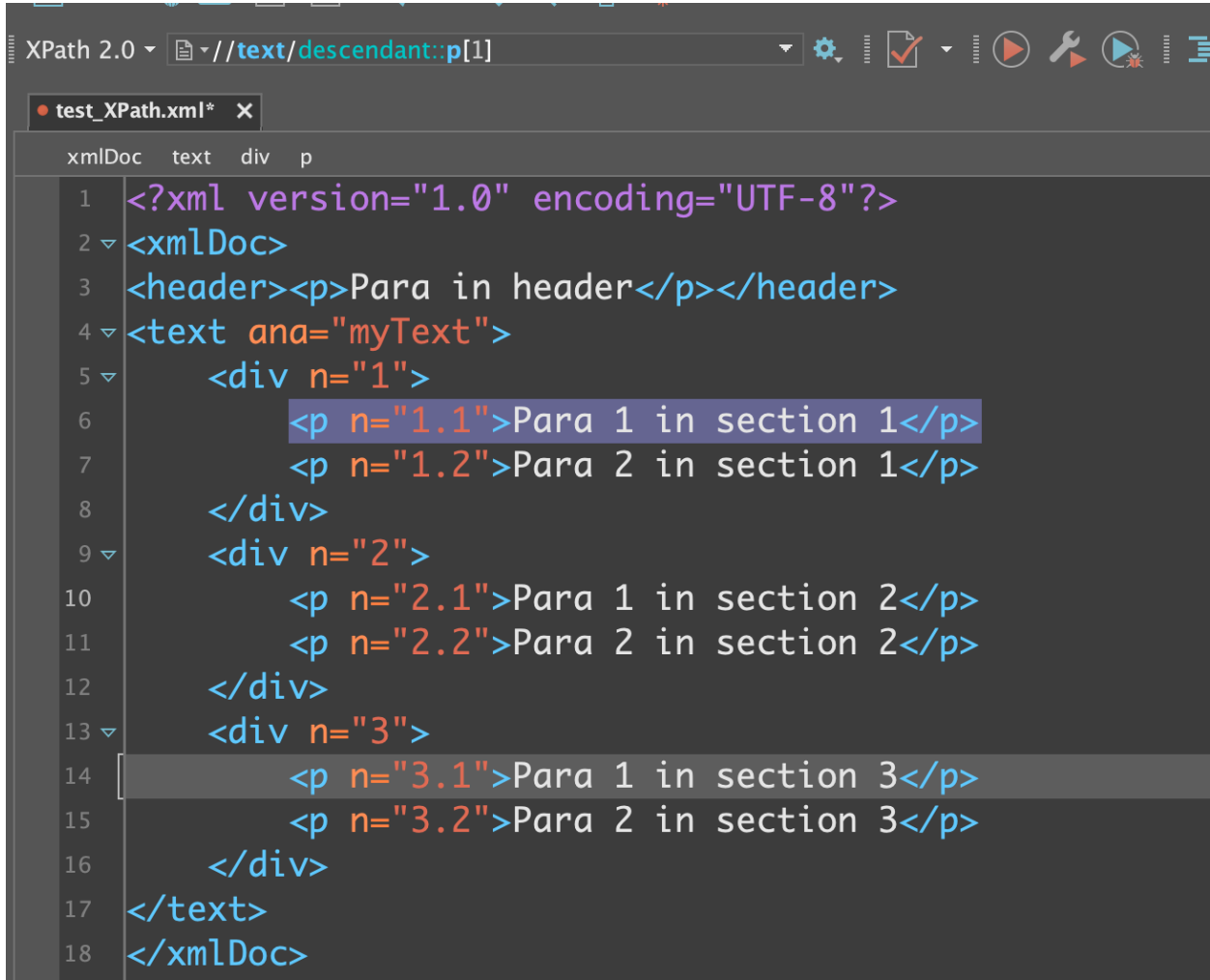
Para 1 in se
Para 2 in se

</text>
</xmlDoc>
Results
Description - 9 items
n="1"
Para 1 in section 1
Para 2 in se
Para 1 in section 1
Press F2 for focus
n="2"
Para 1 in section 2
Para 2 in section 2
n="3"
Para 1 in section 3
Para 2 in section 3
XPath - Dares_Historia_Chap..
XPath - test_XPath.xml

//text//p[1]



//text/descendant::p[1]



Translate to prose
(What does each statement mean?

1. //book[@category='fiction']
2. //*[@type]
3. //chapter[5]/s[1]
4. Will the 3rd example work for the following structure:
   <chapter> <s></s> </chapter>

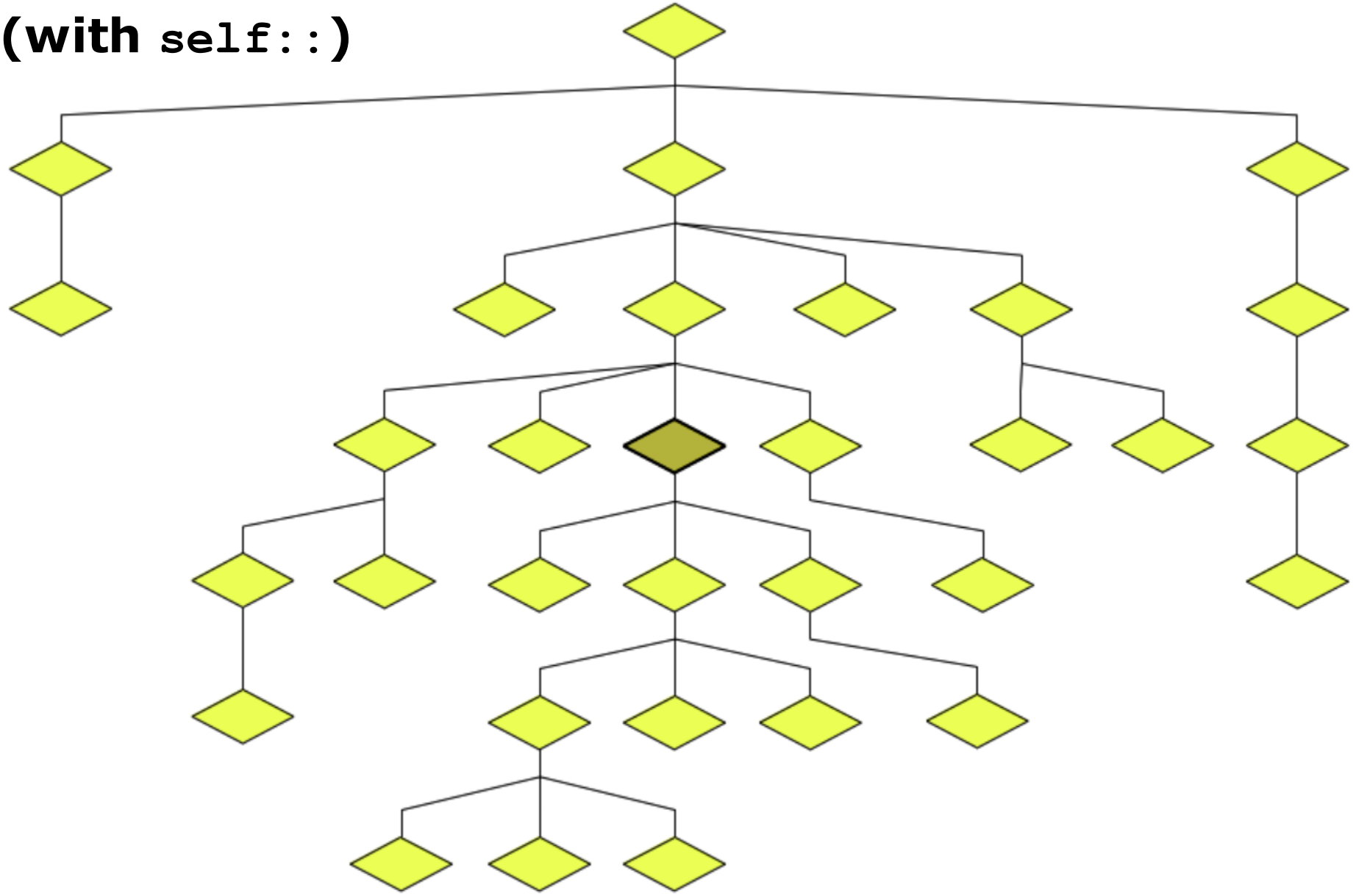
Translate to XPath

1. Give me all items of a list.
2. Give me the first item of a list
3. Give me all elements that have attribute 'ana'
4. Give me all children elements of the first division element, which itself is a child of text
5. Give me all title elements which have an attribute 'type' with the attribute value 'uniform'.

XPath Axes

- An axis represents a relationship to the context (current) node, and is used to locate nodes relative to that node on the tree.

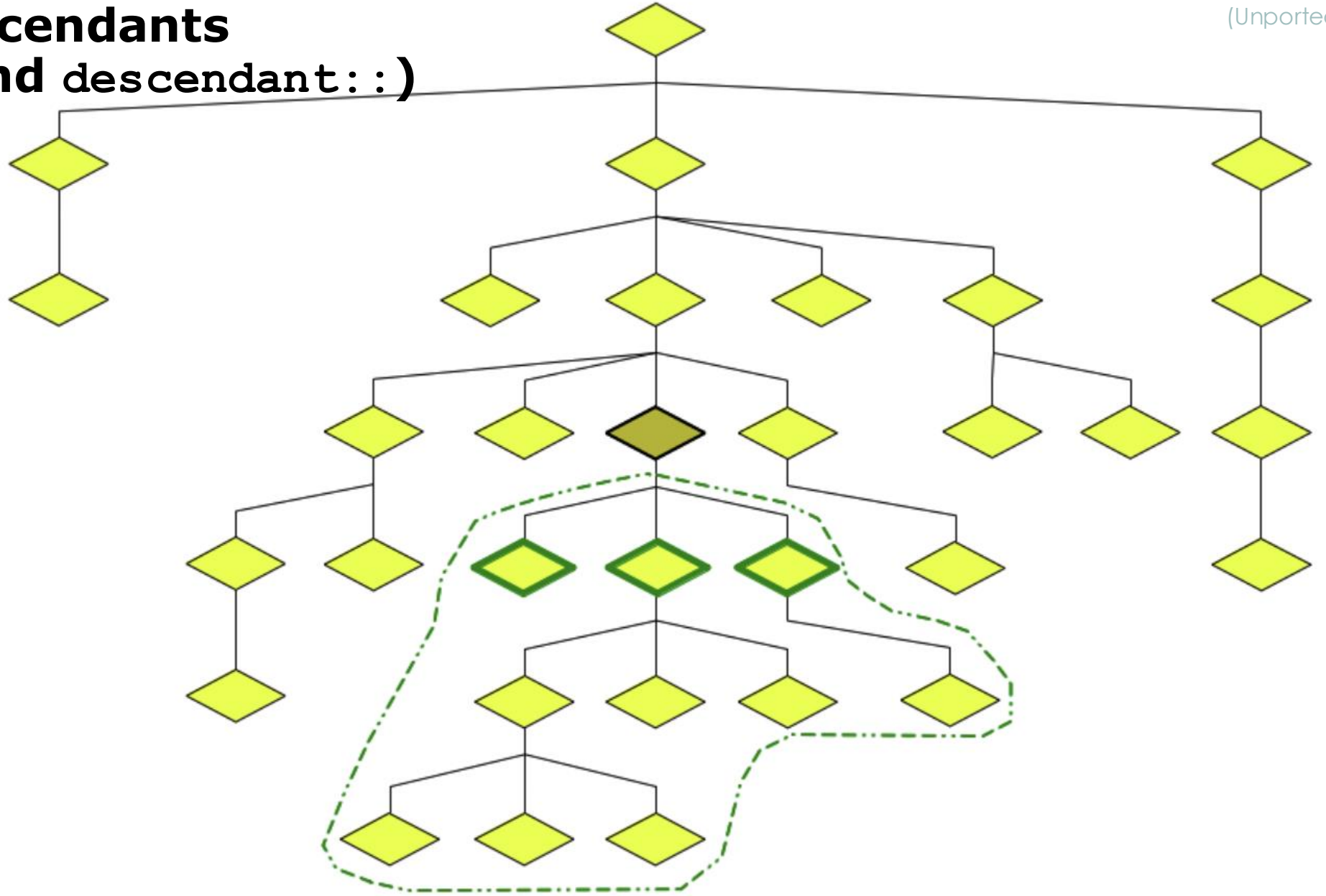
tree (with `self::`)



© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

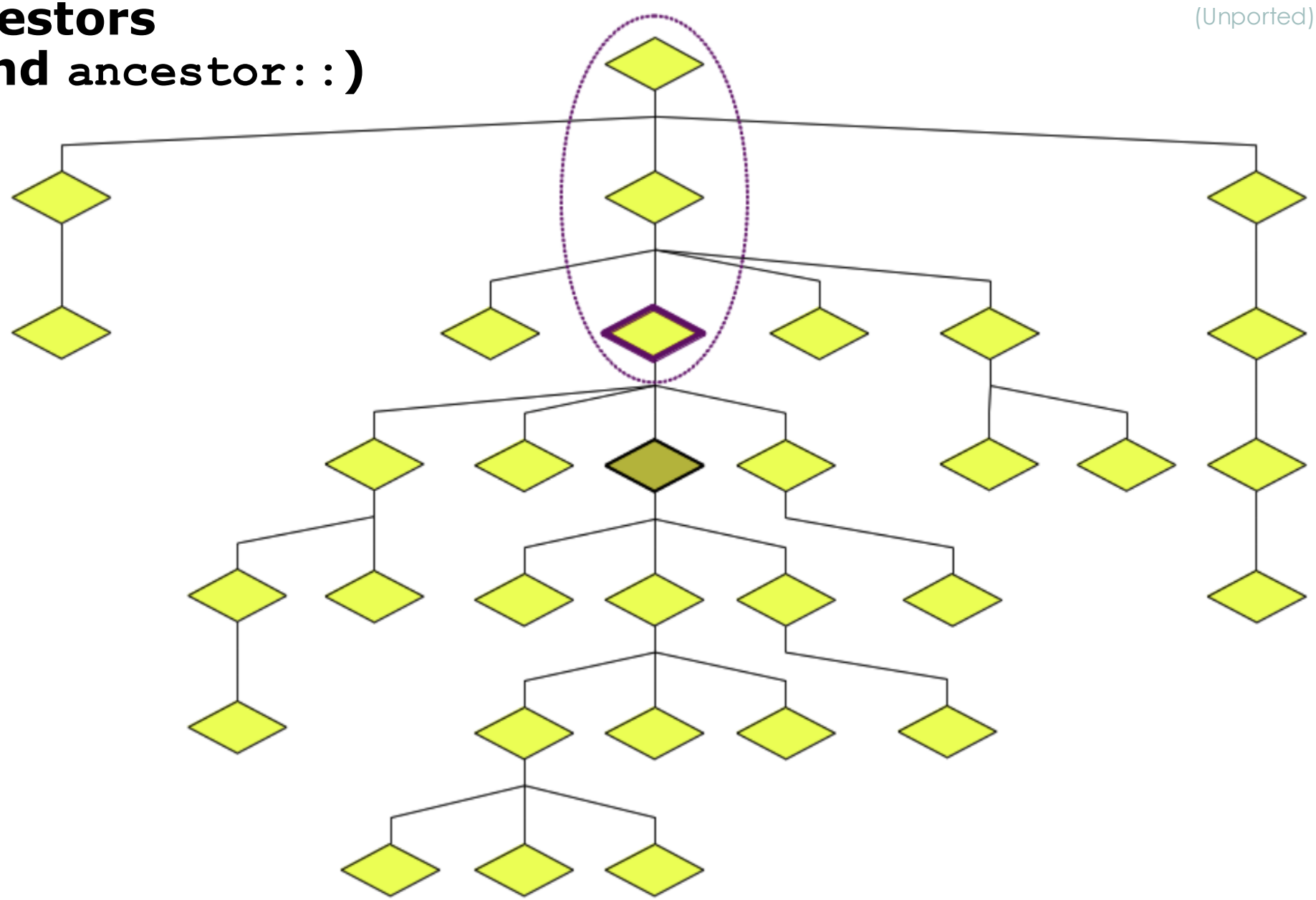
© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

descendants (`child::` and `descendant::`)



© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

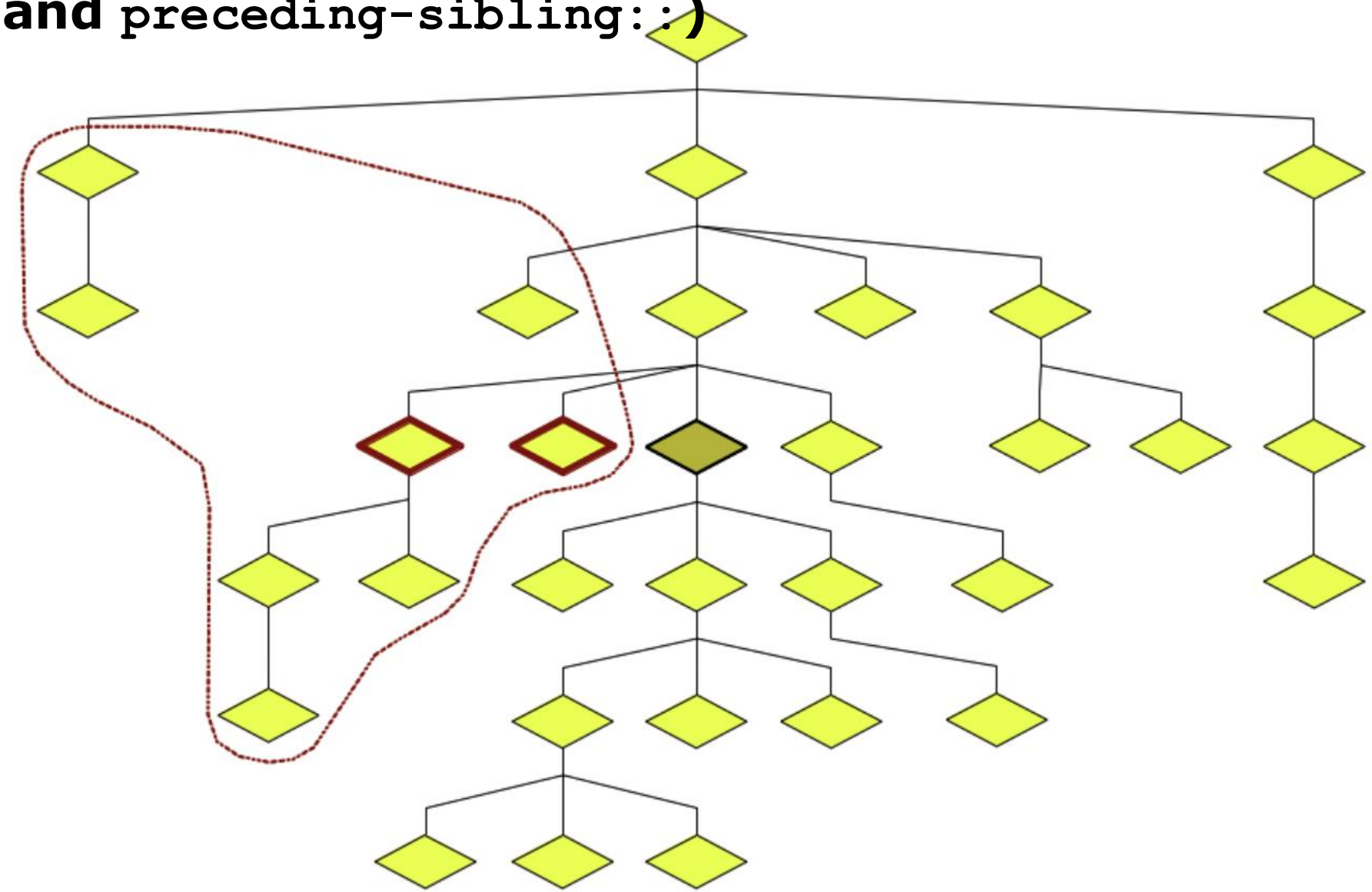
ancestors (`parent::` and `ancestor::`)



© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

preceding

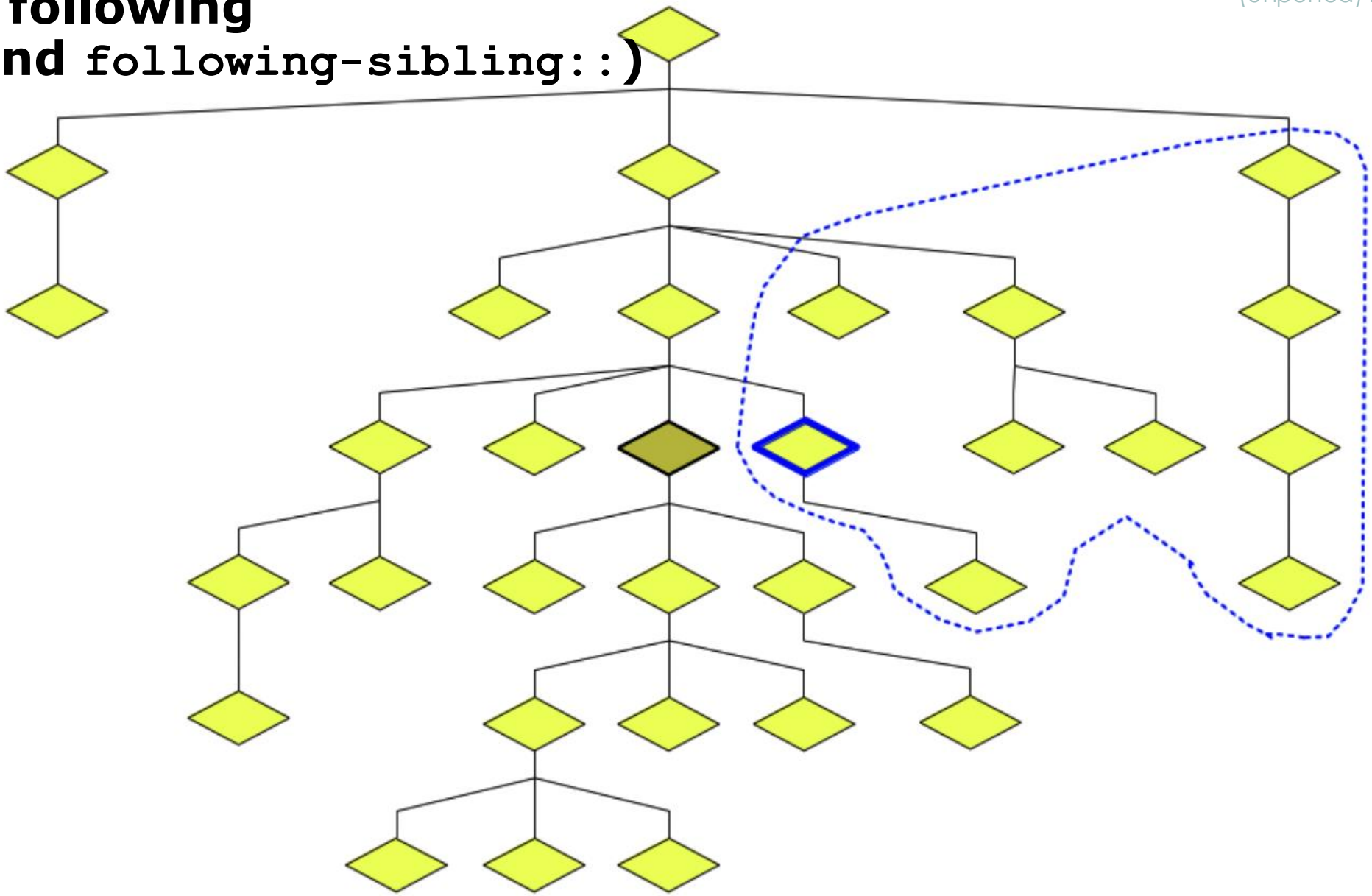
(`preceding::` and `preceding-sibling::`)



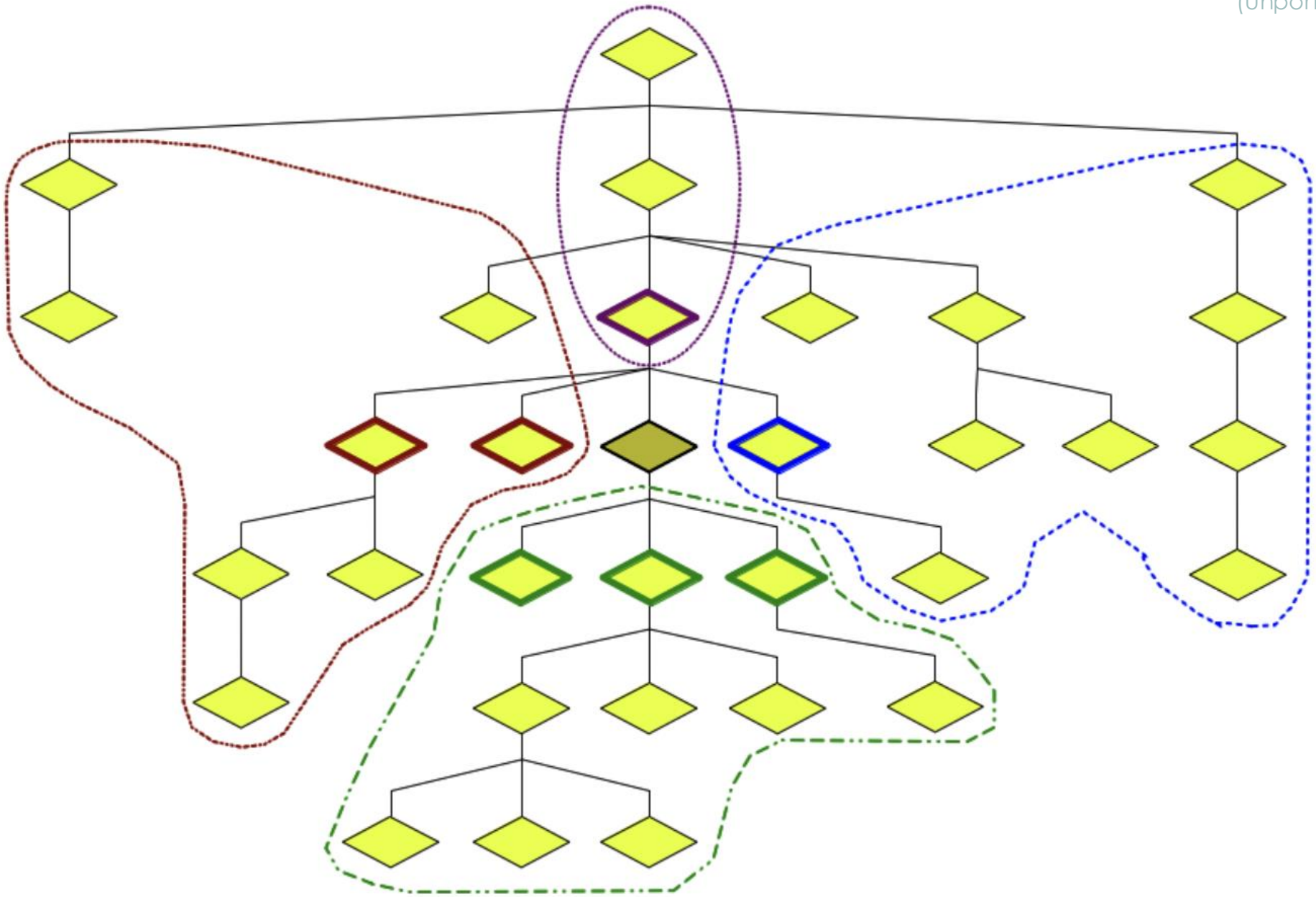
© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.

following

(`following::` and `following-sibling::`)

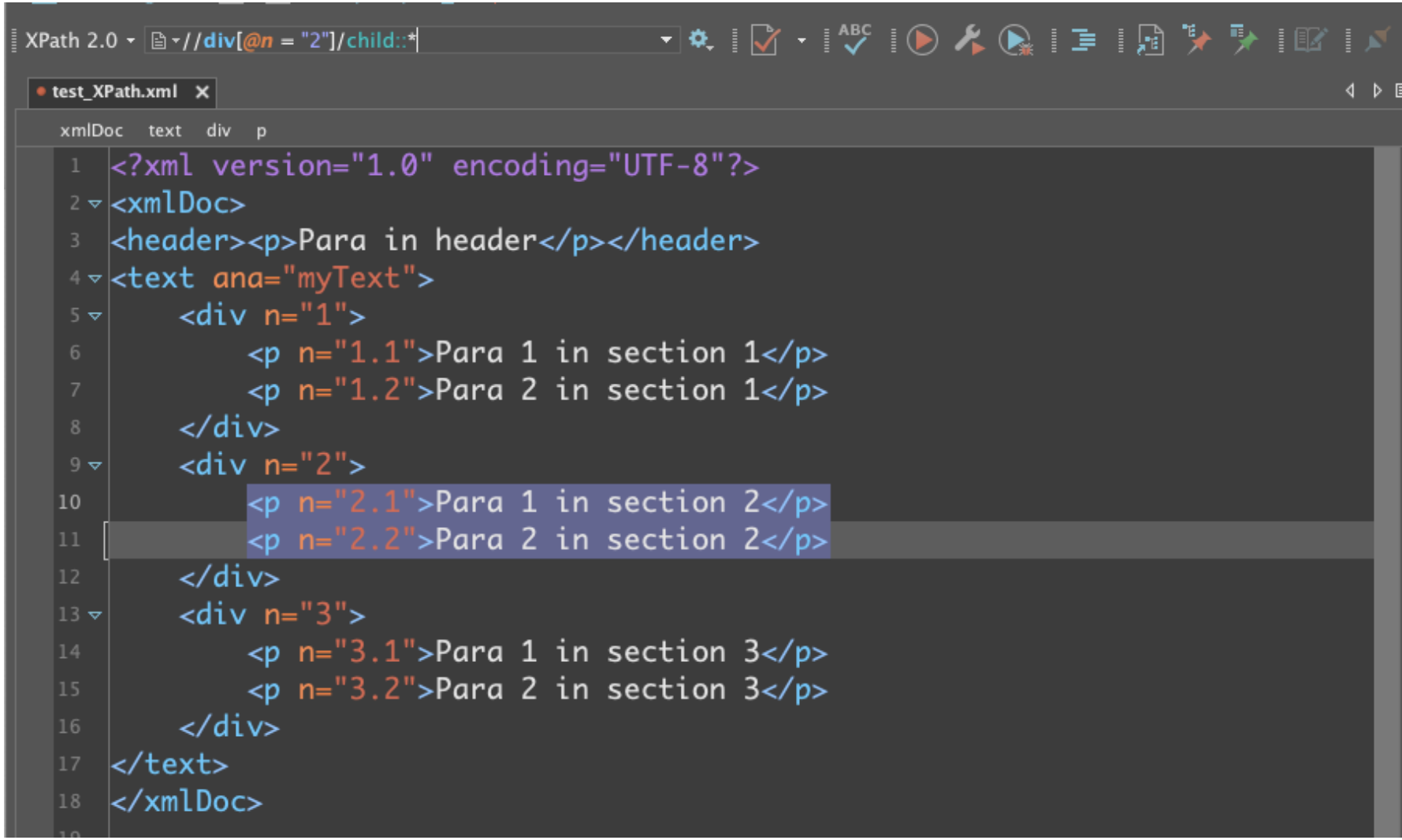


© 2010 Syd Bauman and the Women Writers Project This TEI-encoded XML file is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 (Unported) license.



AxisName	Result
parent	Selects the parent of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
following	Selects everything in the document after the closing tag of the current node
following-sibling	Selects all siblings after the current node
preceding	Selects all nodes that appear before the current node in the document, except ancestors, attribute nodes and namespace nodes
preceding-sibling	Selects all siblings before the current node

//div[@n = "2"]/child::*


XPath 2.0
//div[@n = "2"]/child::*
test_XPath.xml
xmlDoc text div p
<?xml version="1.0" encoding="UTF-8"?>
<xmlDoc>
<header> Para in header </header>
<text ana="myText">

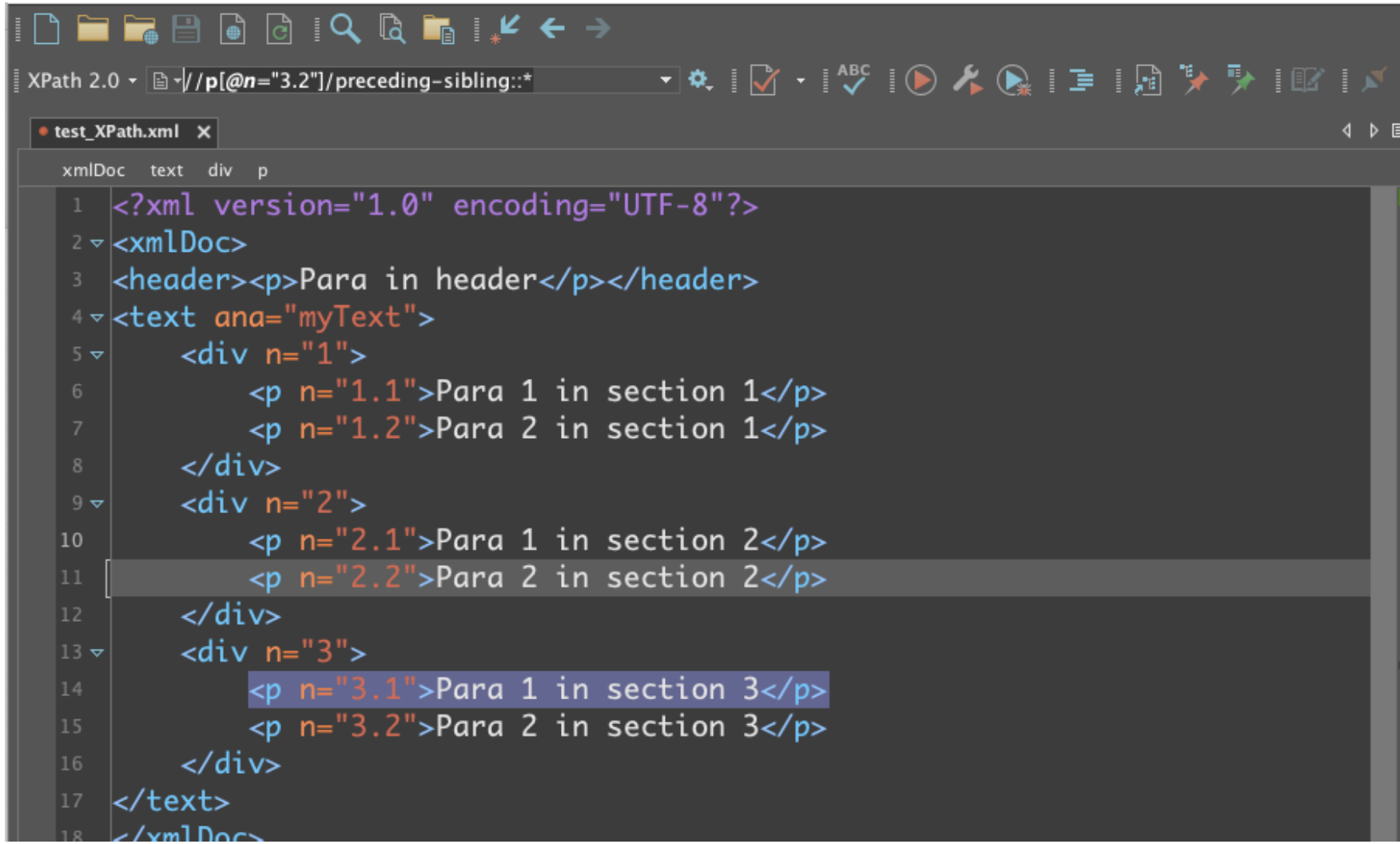
Para 1 in section 1
Para 2 in section 1

Para 1 in section 2
Para 2 in section 2

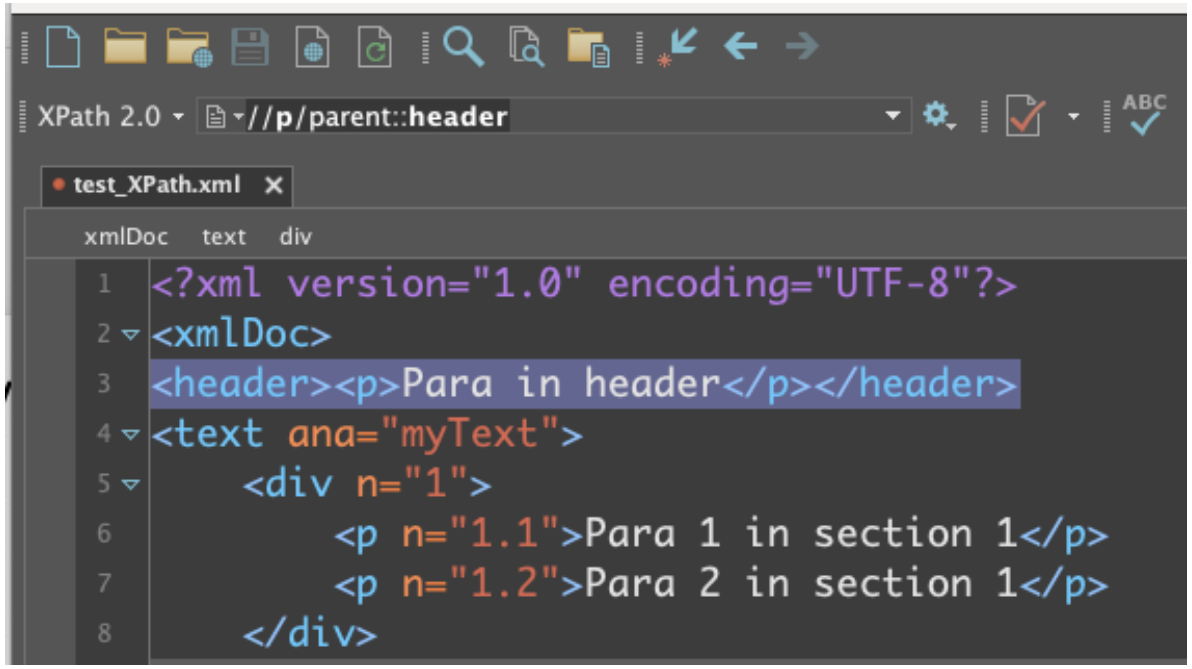
Para 1 in section 3
Para 2 in section 3

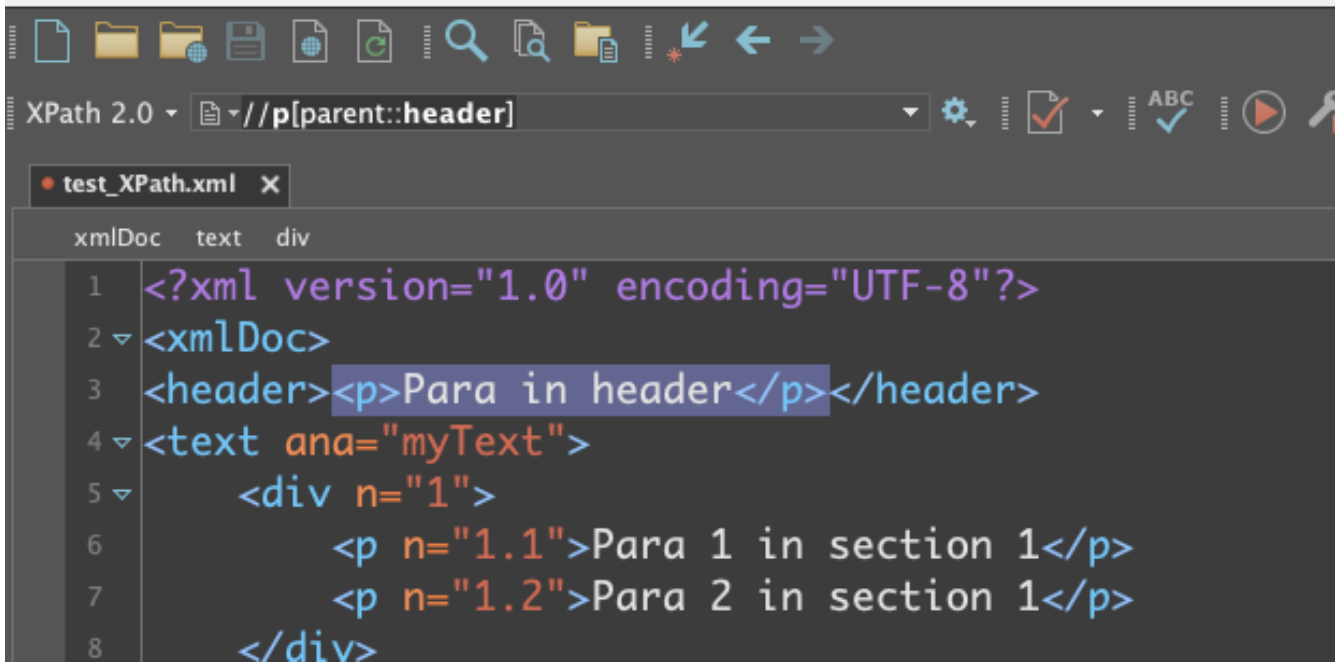
</text>
</xmlDoc>

//p[@n="3.2"]/preceding-sibling::*



Be careful with your predicates & axes!





//p/parent::header ≠ //p[parent::header]

Exercise 2

Using ***Dares_annotations_spoiler.xml***:

1. Find all **geographical** coordinates in the **authority list.** How many did you find?
2. Find all names of **places** mentioned in the **text.** How many did you find?
3. Find out which **place name** is the **first** one mentioned in **the third sentence.**
4. Find out which **person name** is the fourth one mentioned **in the chapter**.
5. Using the **preceding-sibling axis** find the name of the person mentioned in the same sentence as Podacres (Pod_001)

Exercise 3.1

- **In groups:**
  - Prepare **five** XPath questions (in prose) for other groups to answer for one of the following files.
    - **Group 1:** Munich_Clm_305_transcr_spoiler.xml
    - **Group 2:** Dares_edition_spoiler.xml
    - **Group 3:** Paris_Latin_5691_description_spoiler.xml
    - **Group 4:** Bodleian_FirstFolio_spoiler.xml
  - Prepare answers to these questions (XPath & Result)

Exercise 3.2

- Add **only** your questions to this doc:
- https://tinyurl.com/TNAH2025Quiz