Modelling humanities data with TEI-XML

SCHOLARLY EDITING AND MANUSCRIPT CATALOGUING IN THE DIGITAL AGE

Dr Katarzyna Anna Kapitan
20 November 2024

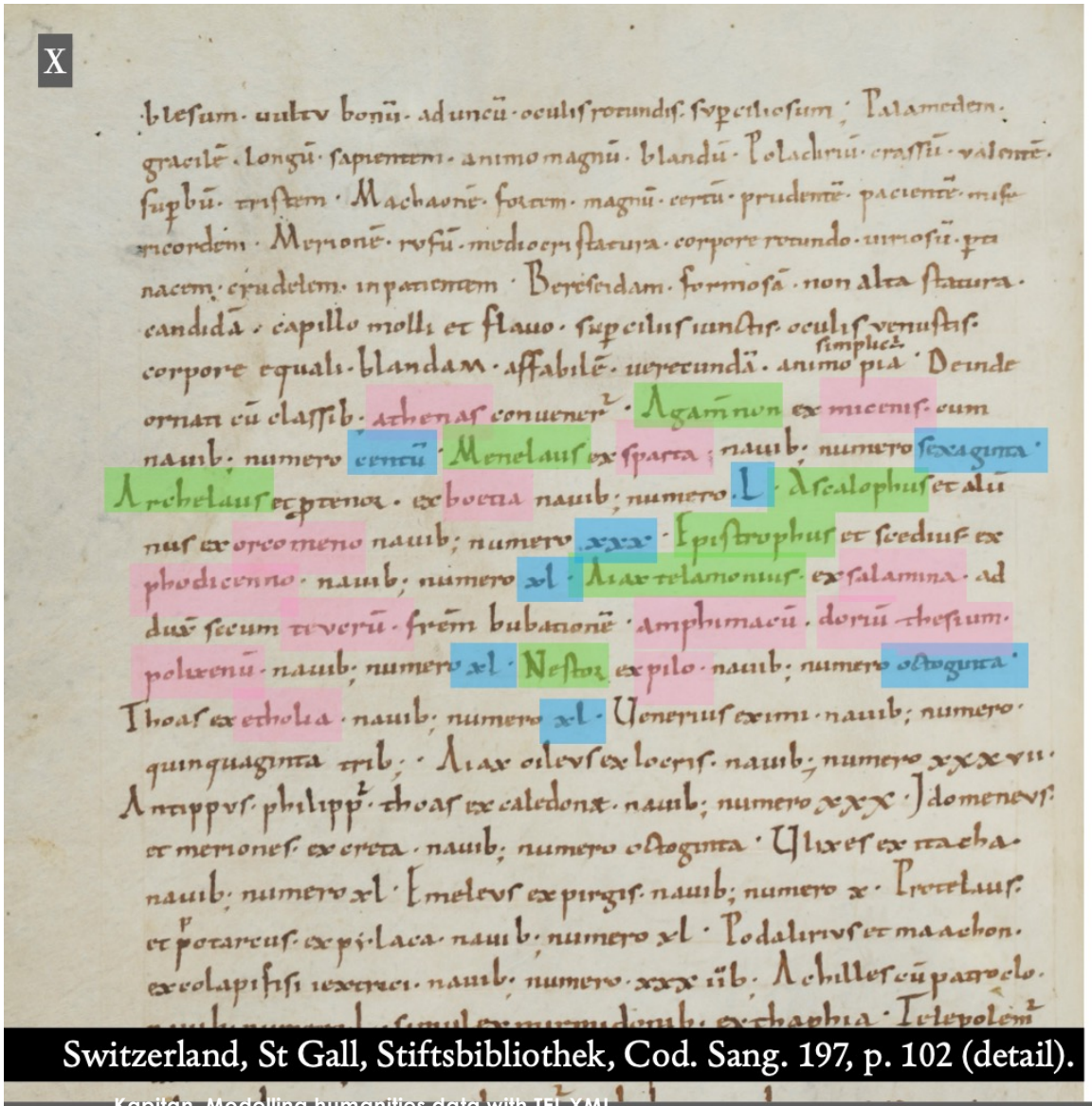
Annotations Recap

There were 44 leaders sailing from Greece to Troy. Their names are: Agamemnon Menelaus Arcesilaus Prothoenor Ascalaphus Ialmenusus Epistrophus Schedius Aiax Telamonius Teucrus Amphimachus Diores Thalpius Polyxenus Nestor Thoas Nierus Aiax Oileus Antiphus Phidippus Idomeneus Meriones Ulixes Eumelus Protesilaus Podacres Podalirius Machaon Achilles Patroclo Tlepolemus Eurypylus Antiphus Amphimachus Polypoetes Leonteus Diomedes Euryalus Sthenelus Philoctetes Guneus Prothous Agapenor Menestheus

Their names in alphabetic order are:

- Achilles
- Agamemnon
- Agapenor
- Aiax Oileus
- Aiax Telamonius
- Amphimachus
- Amphimachus
- Antiphus
- Antiphus

```
            <listPerson>
               <person xml:id="Aga_001">
                  <persName>Agamemnon</persName>
                  <idno type="wikidata">Q128176</idno>
                  <note>King of Mycenae; son of Atreus in Greek mythology.</note>
               </person>
            </listPerson>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <div type="chapter" n="14">
            <p>
               <s>Deinde ornati cum classe Graeci Athenas convenerunt:</s>
               <s><persName ref="#Aga_001">Agamemnon</persName> ex Mycenis cum navibus numero
```

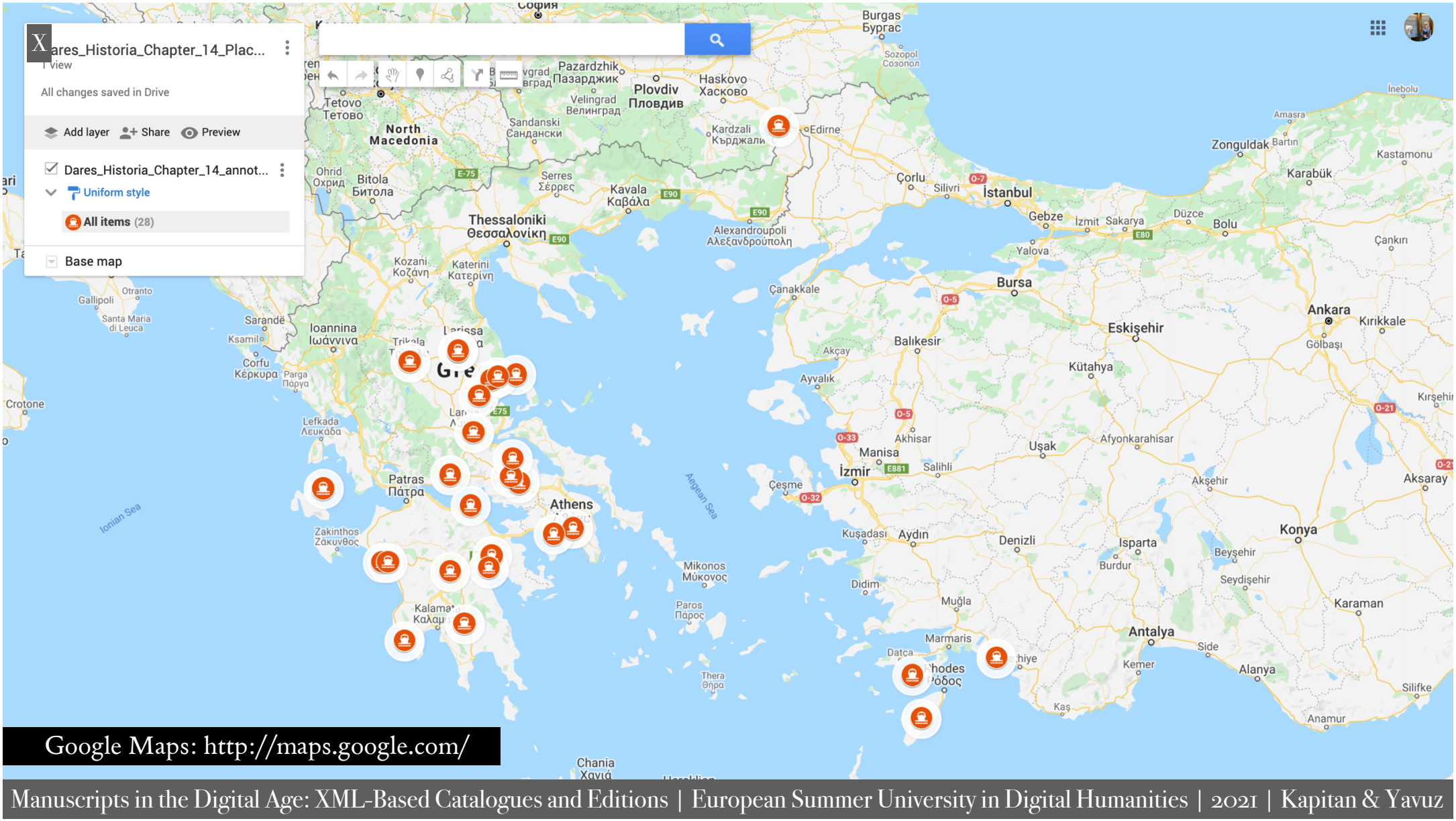


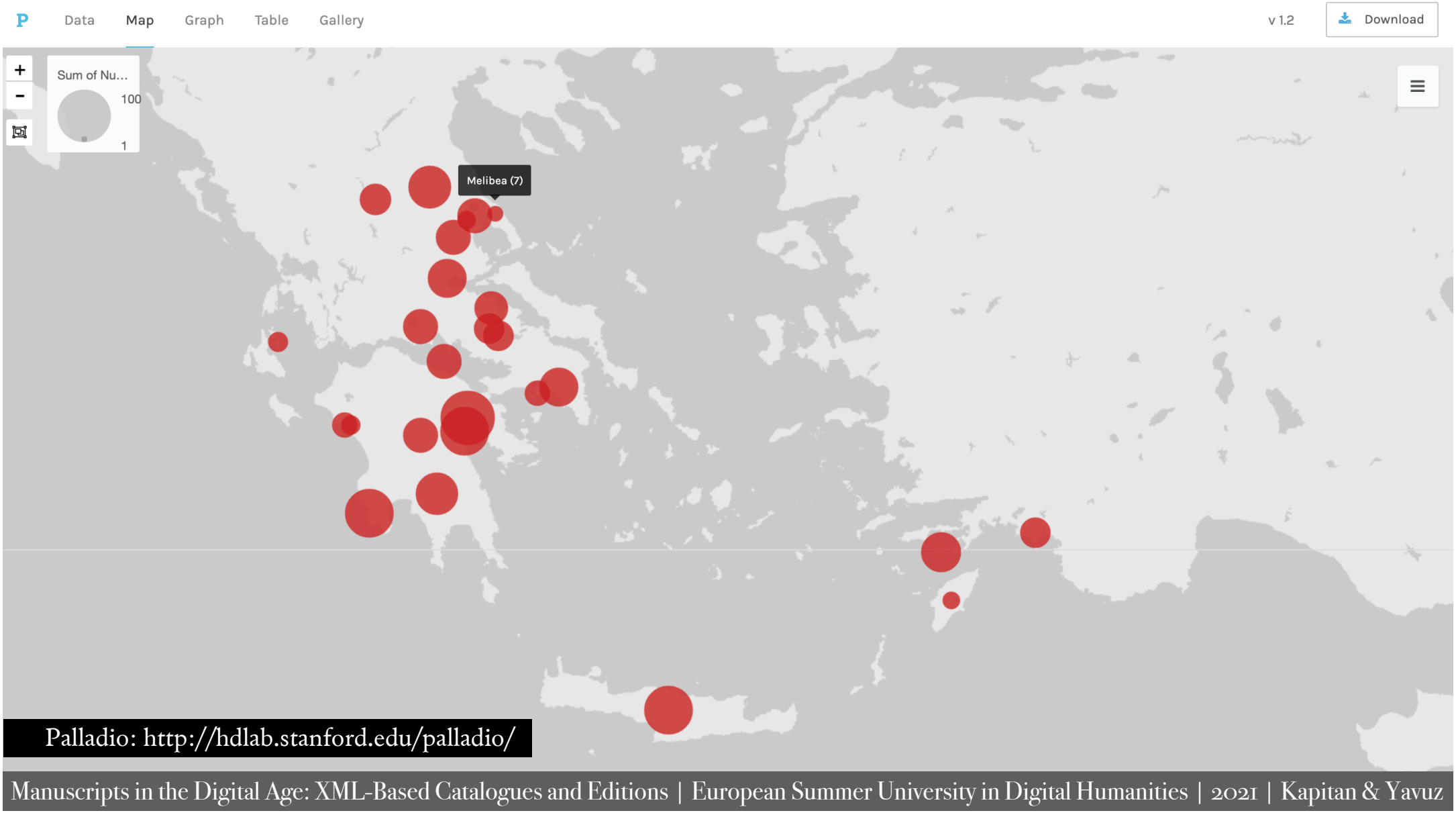
Switzerland, St Gall, Stiftsbibliothek, Cod. Sang. 197, p. 102 (detail).

There were 44 leaders sailing from Greece to Troy. Their names are: Agamemnon Amphimachus Diores Thalpius Polyxenus Nestor Thoas Nierus Aiax Oileus Antip Tlepolemus Eurypylus Antiphus Amphimachus Polypoetes Leonteus Diomedes Eu

Their names in alphabetic order are:

- Achilles
- Agamemnon
- Agapenor
- Aiax Oileus
- Aiax Telamonius
- Amphimachus
- Amphimachus
- Antiphus
- Antiphus
- Arcesilaus
- Ascalaphus
- Diomedes
- Diores
- Epistrophus
- Eumelus
- Euryalus
- Eurypylus
- Guneus
- Ialmenusus
- Idomeneus
- Leonteus
- Machaon
- Menelaus
- Menestheus
- Meriones
- Nestor
- Nierus
- Patroclo
- Phidippus
- Philoctetes
- Podacres


ares_Historia_Chapter_14_Plac...
1 view
All changes saved in Drive
Add layer
Share
Preview
Dares_Historia_Chapter_14_annot...
Uniform style
All items (28)
Base map
Google Maps: http://maps.google.com/


P
Data
Map
Graph
Table
Gallery
v 1.2
Download
Sum of Nu...
100
1
Melibea (7)
Palladio: http://hdlab.stanford.edu/palladio/

XPath

What is XPath?

XPath is a major element in the XSLT standard.

XPath can be used to navigate through elements and attributes in an XML document.

- XPath stands for XML Path Language
- XPath uses "path like" syntax to identify and navigate nodes in an XML document
- XPath contains over 200 built-in functions
- XPath is a major element in the XSLT standard
- XPath is a W3C recommendation

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

XPath

XPath – Important for, among other things:
Data export, transformations (XSLT), finding things in your document…

Below on section from the XSLT file used to extract relevant info from Dares, chapter 14.

```
able name="totalShipNumber" select="sum(text//num/@value)"/>
able name="totalPlaces" select="count(text//placeName)"/>
ame from <xsl:value-of select="$totalPlaces"/> cities in Greece a
ht <xsl:value-of select="$totalShipNumber"/> ships.</p>
city brought the following number of ships: <ul>
xsl:for-each select="text//s/placeName">
```

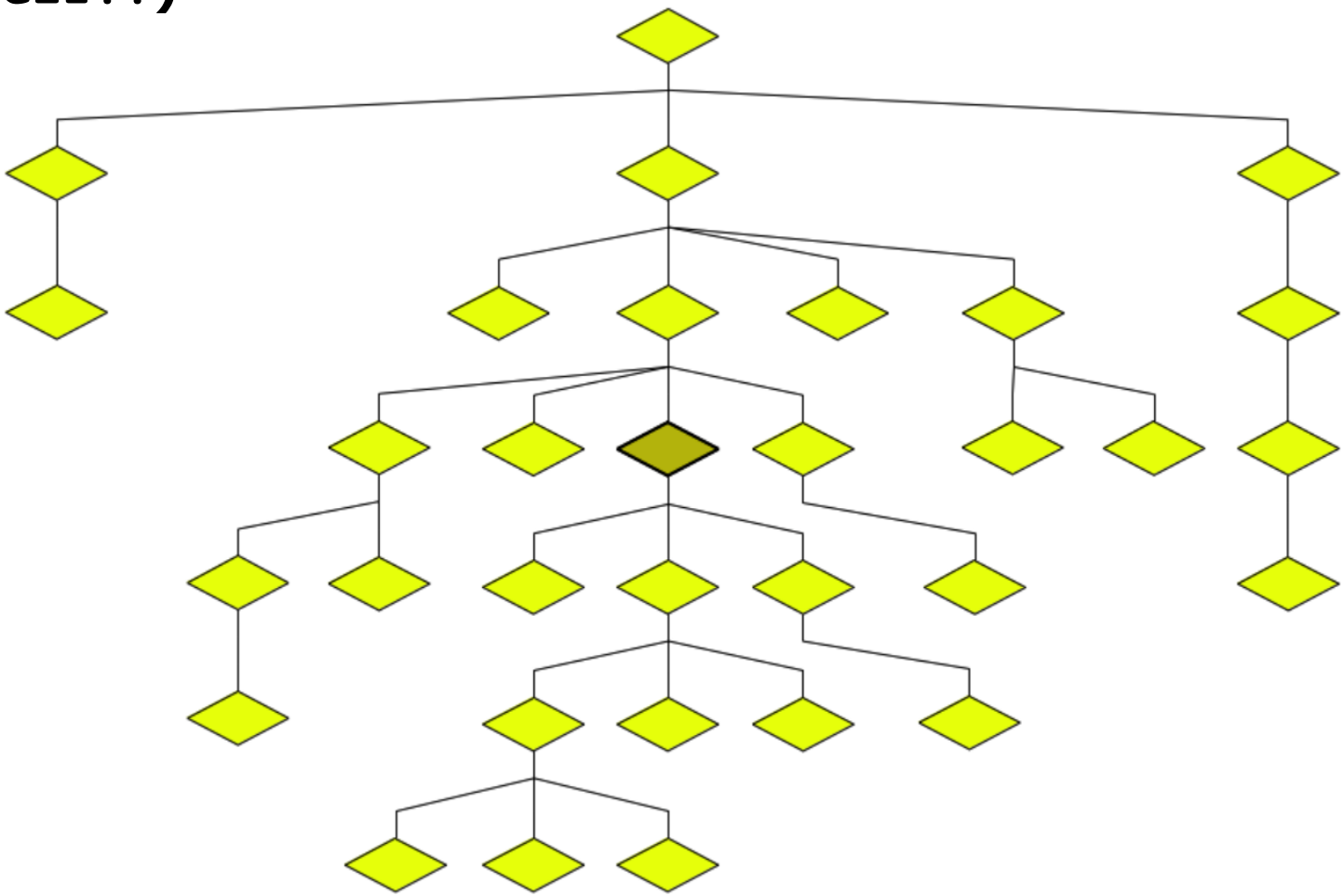
Basic filepath-like path expressions

A bare-bones path expression is similar to filesystem addressing: if the path starts with a forward slash ("/"), then it represents a path from the root; if it does not start with a solidus then it represents a path from "here"

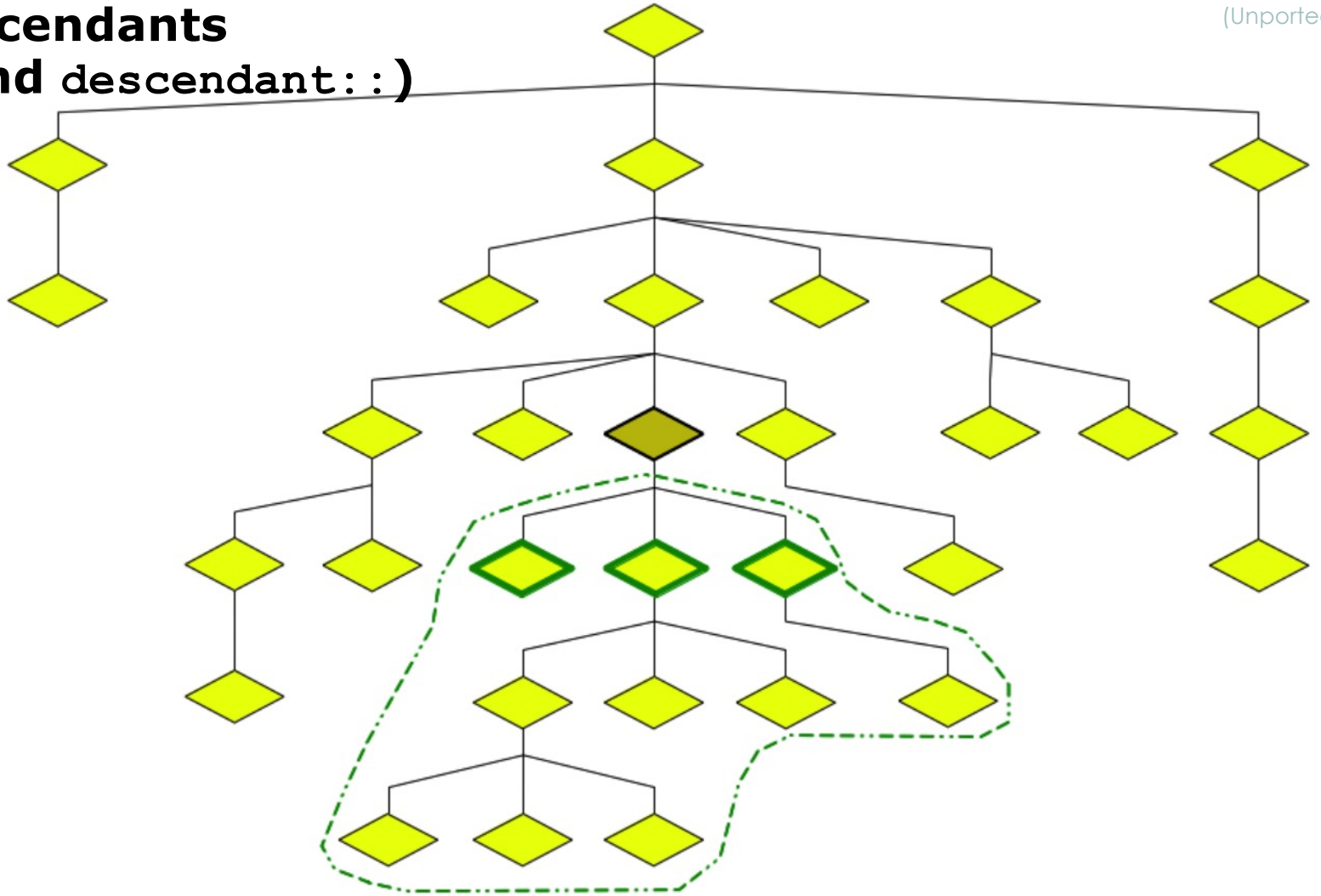
```
/TEI/teiHeader/fileDesc/titleStmt/title
```

```
list/item/label
```

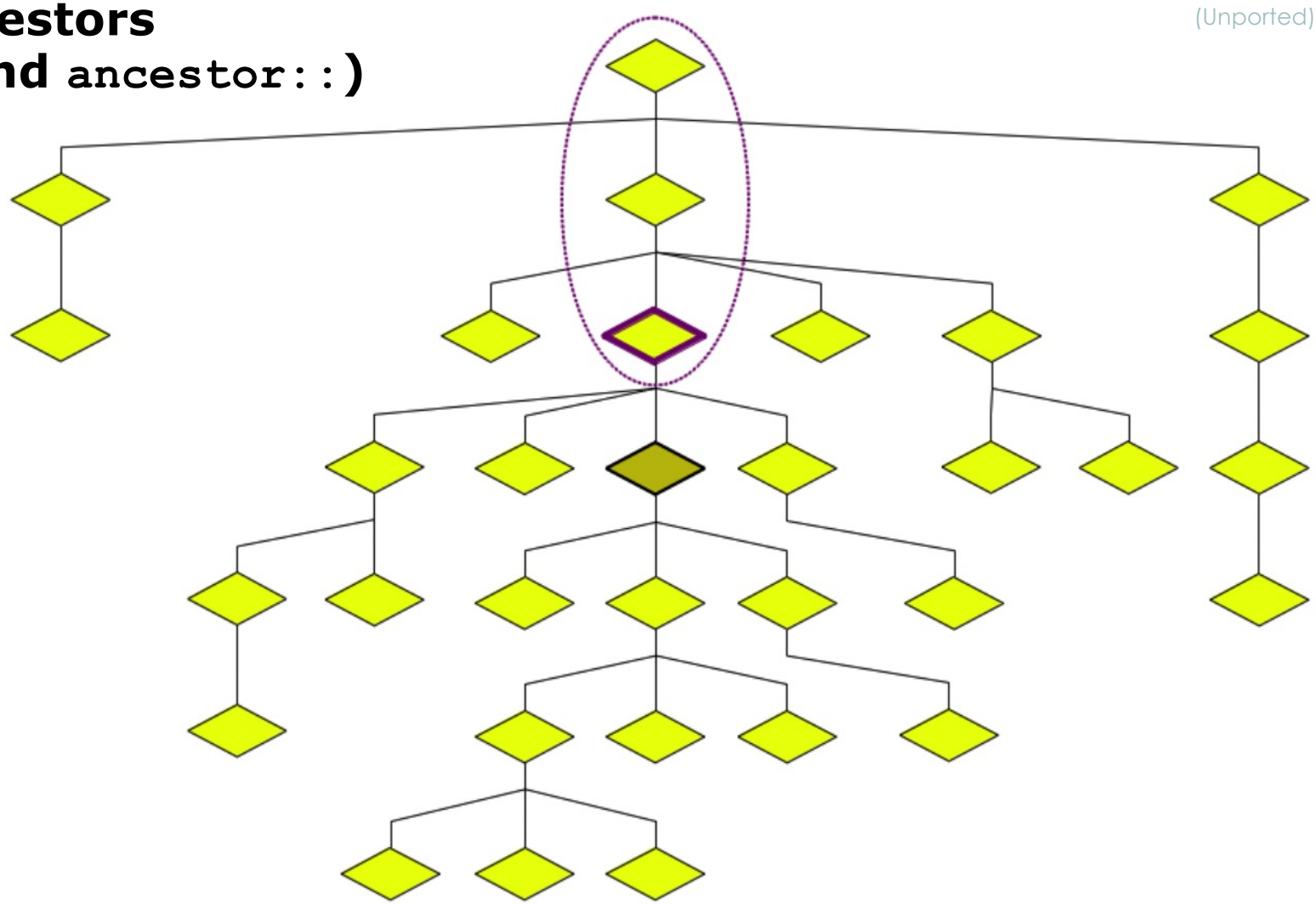
tree (with `self::`)



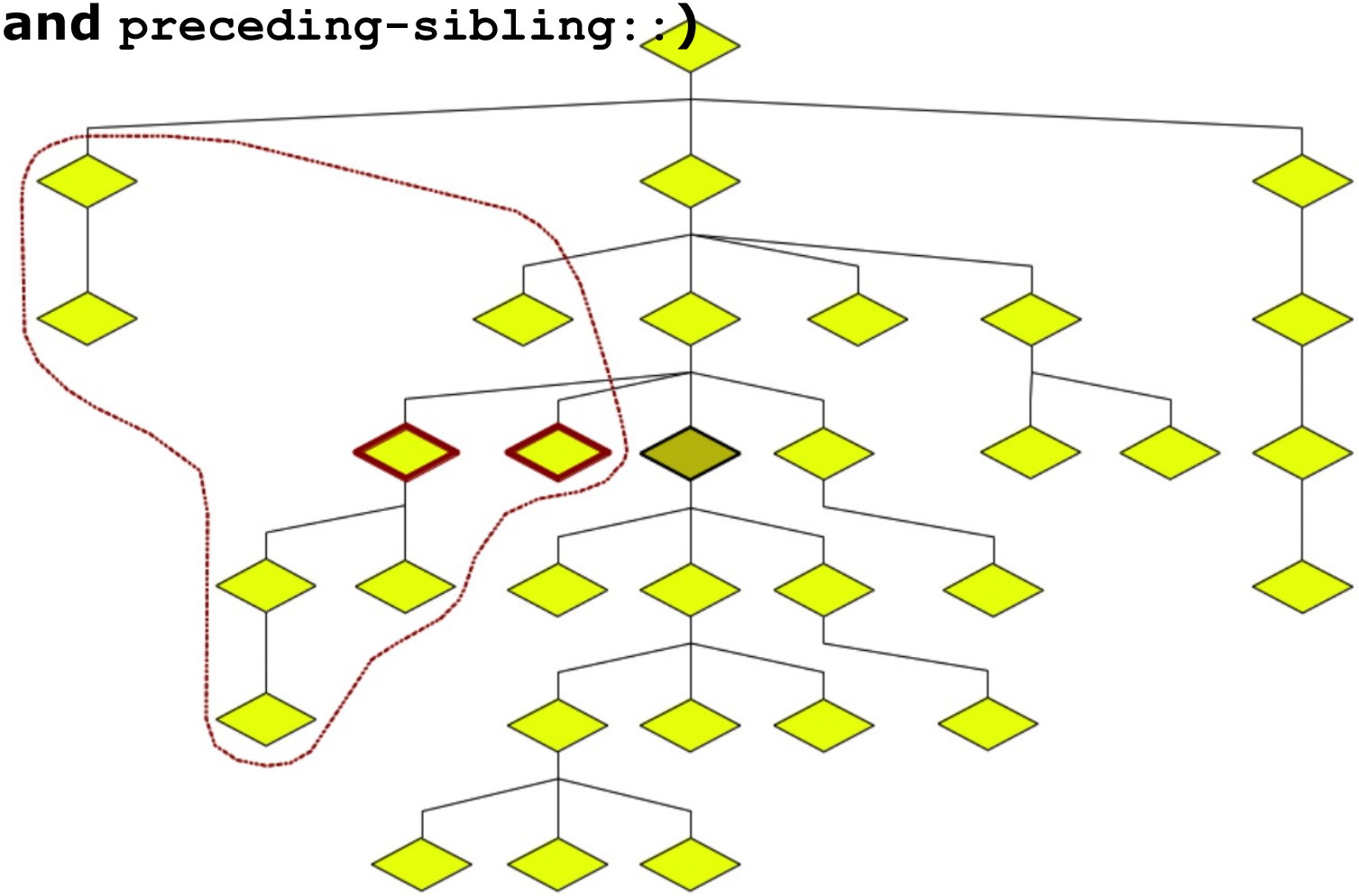
descendants (`child::` and `descendant::`)



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

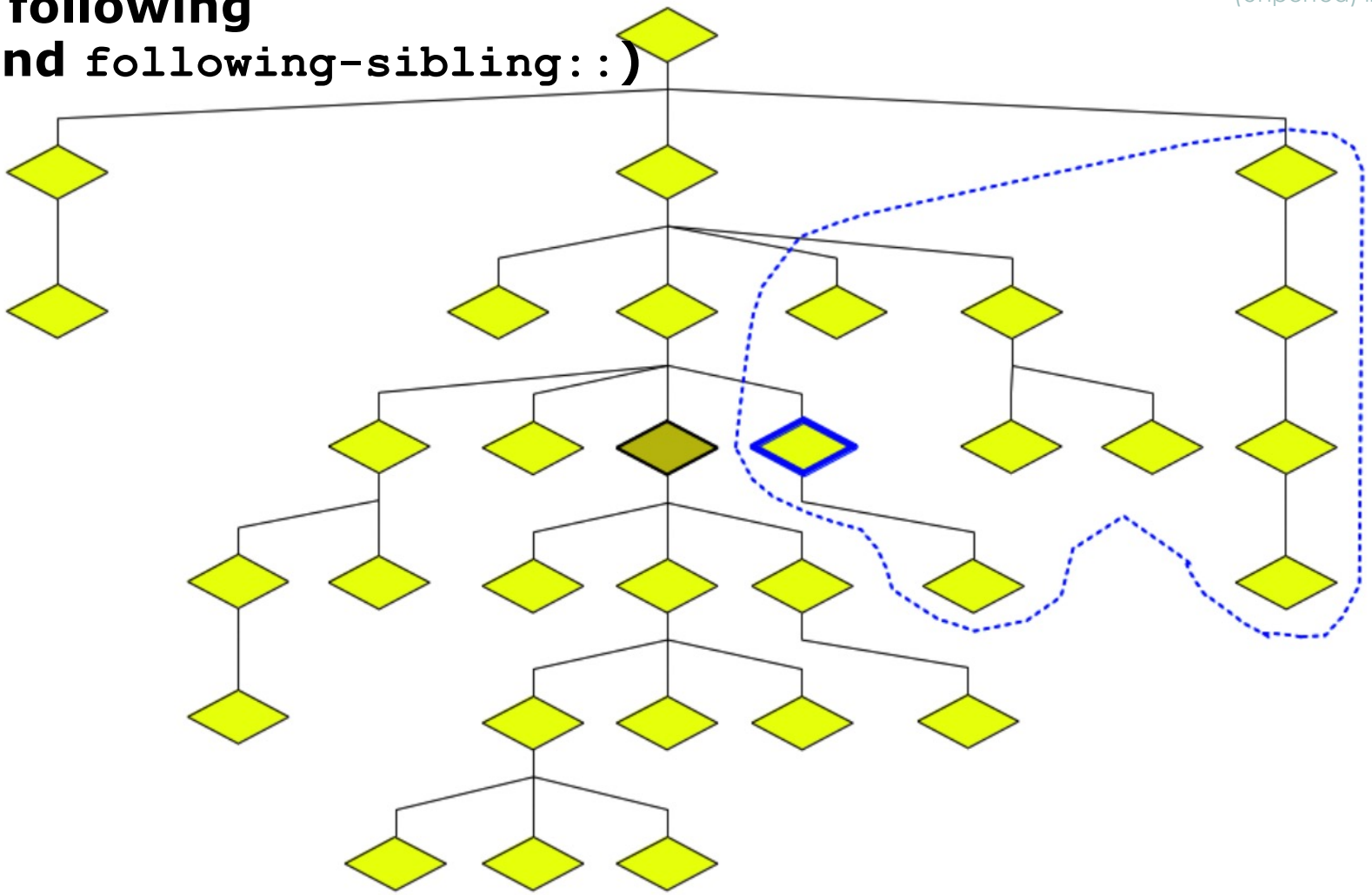


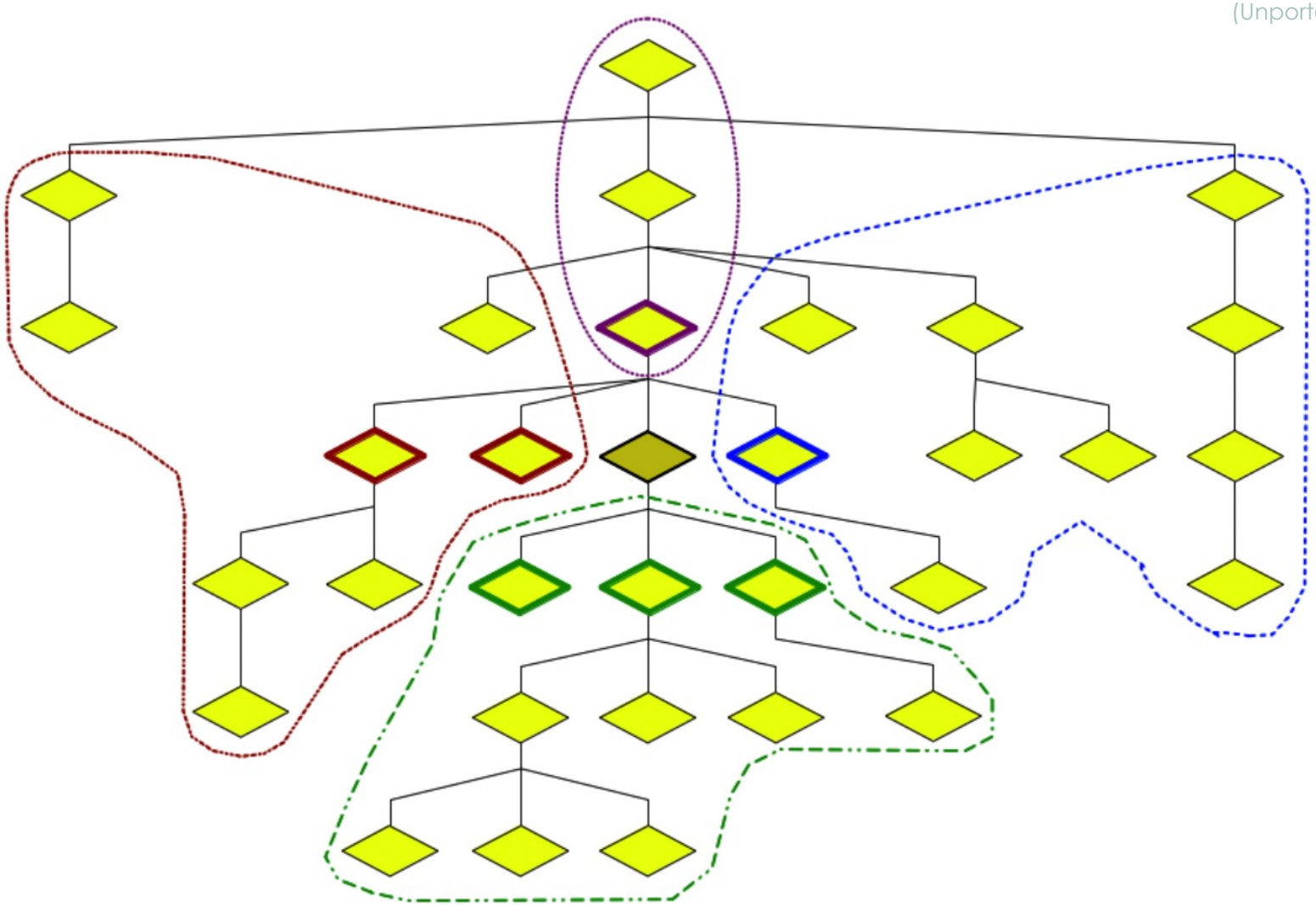
preceding (`preceding::` and `preceding-sibling::`)



following

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





XPath Syntax

Expression	Description
nodename	Selects all nodes with the name "*nodename*"
/	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

XPath Syntax

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

Path Expression	Result
bookstore	Selects all nodes with the name "bookstore"
/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

AxisName	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself
attribute	Selects all attributes of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself
following	Selects everything in the document after the closing tag of the current node

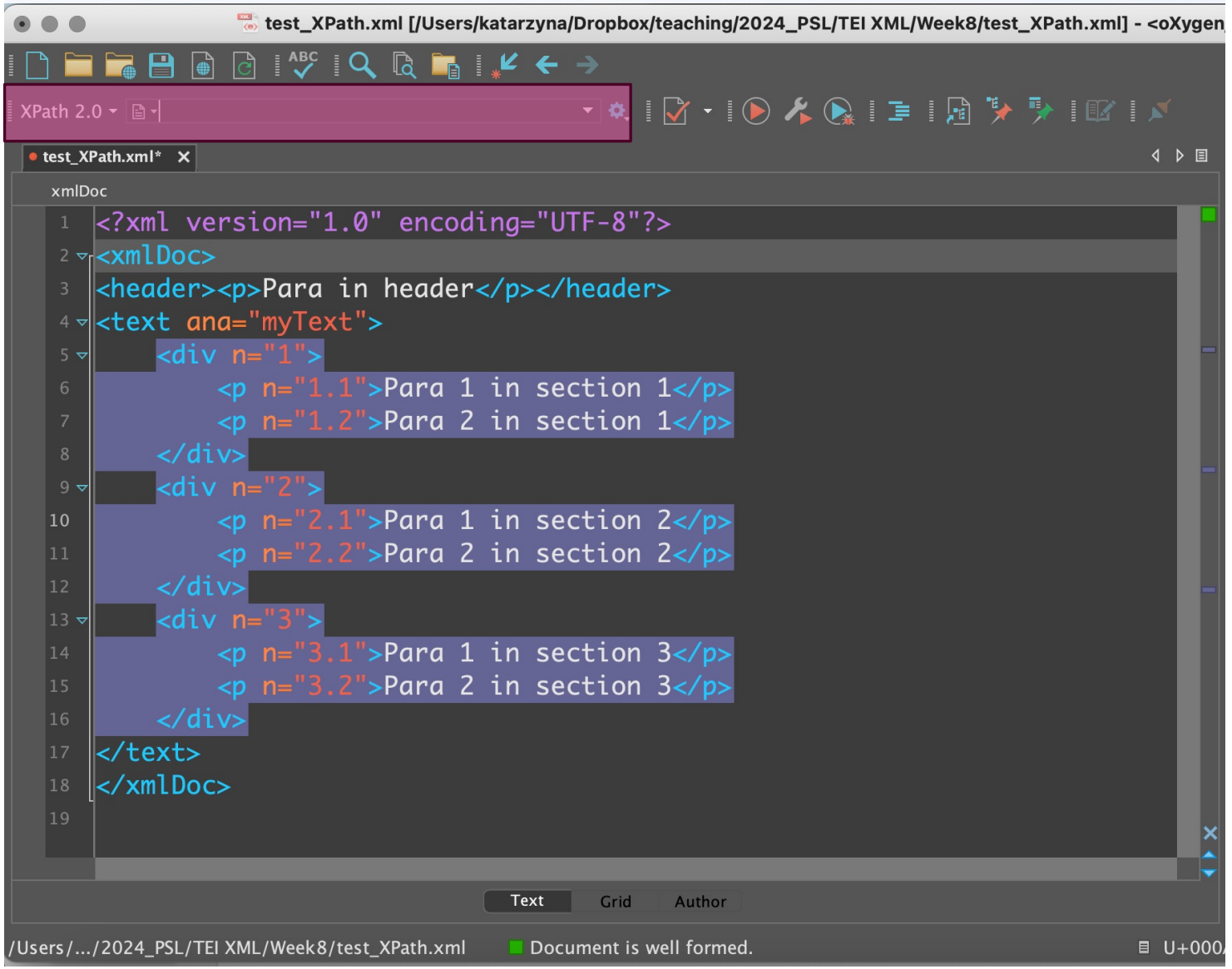
https://www.w3schools.com/xml/xpath_axes.asp

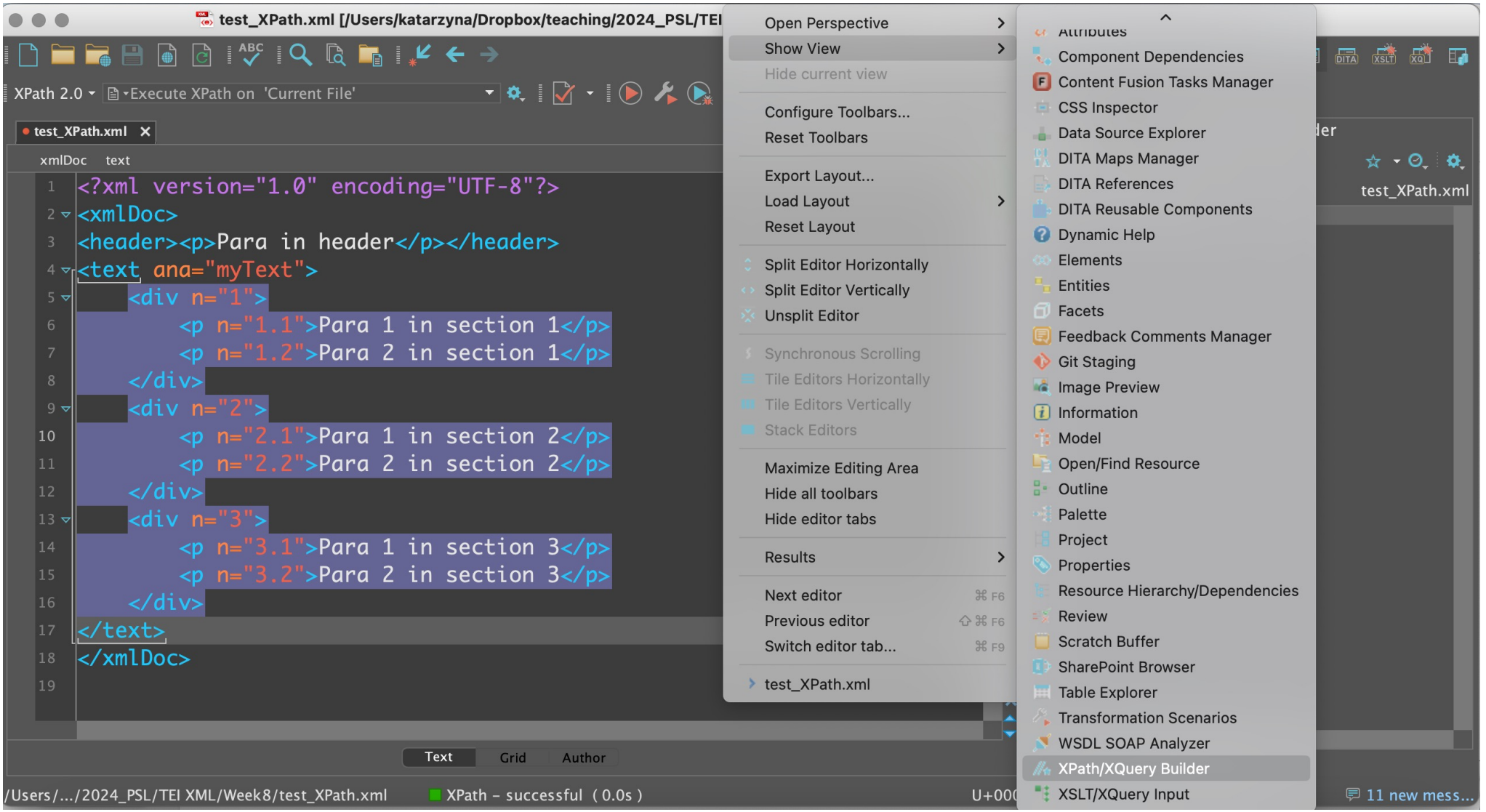
Testing XPath in Oxygen

test_Xpath.xml

On GitHub: /Classes/Week8/Exercises/test_XPath.xml

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.
U+000


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
<?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
```

//parent::header

```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>
```

//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

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

Exercise 1

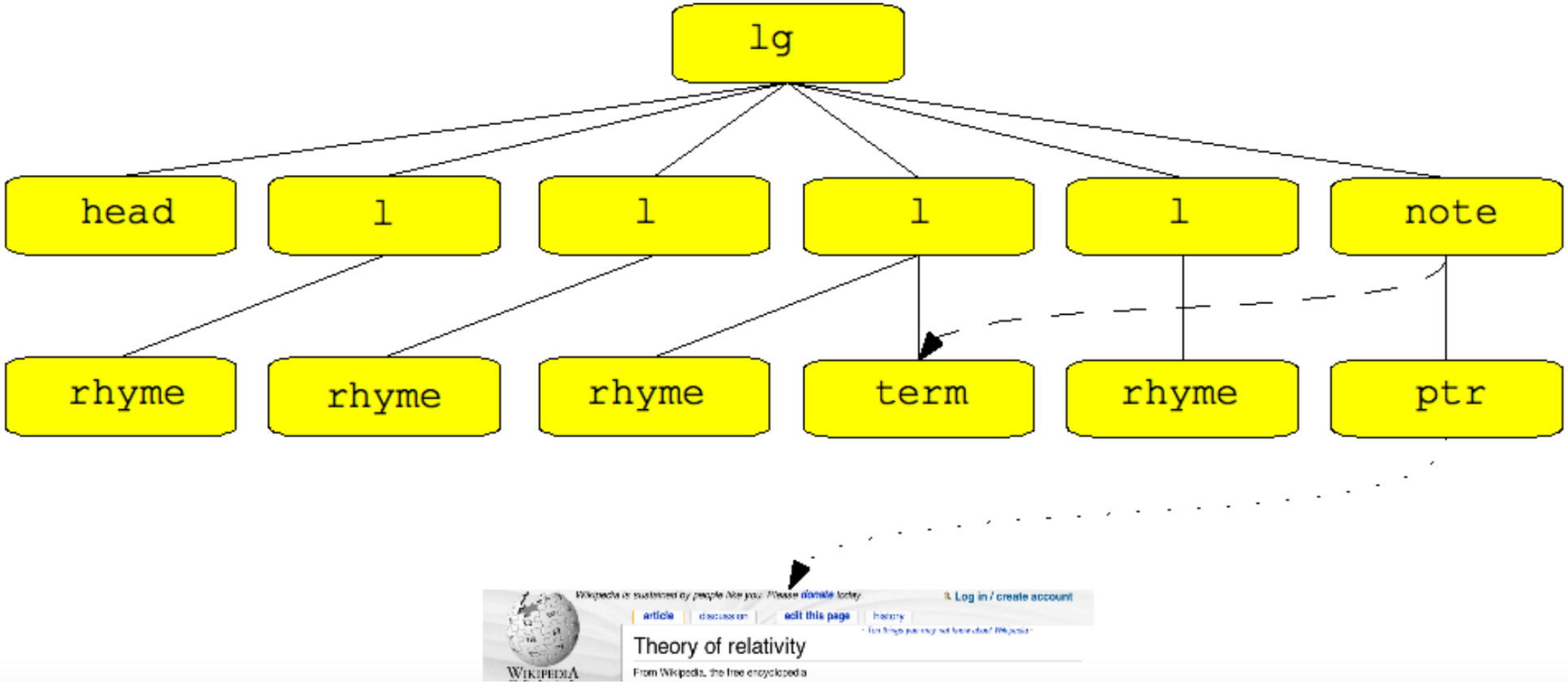
Using Dares_Historia_Chapter_14_annotations_spoiler.xml (Week8/Exercises), find all:

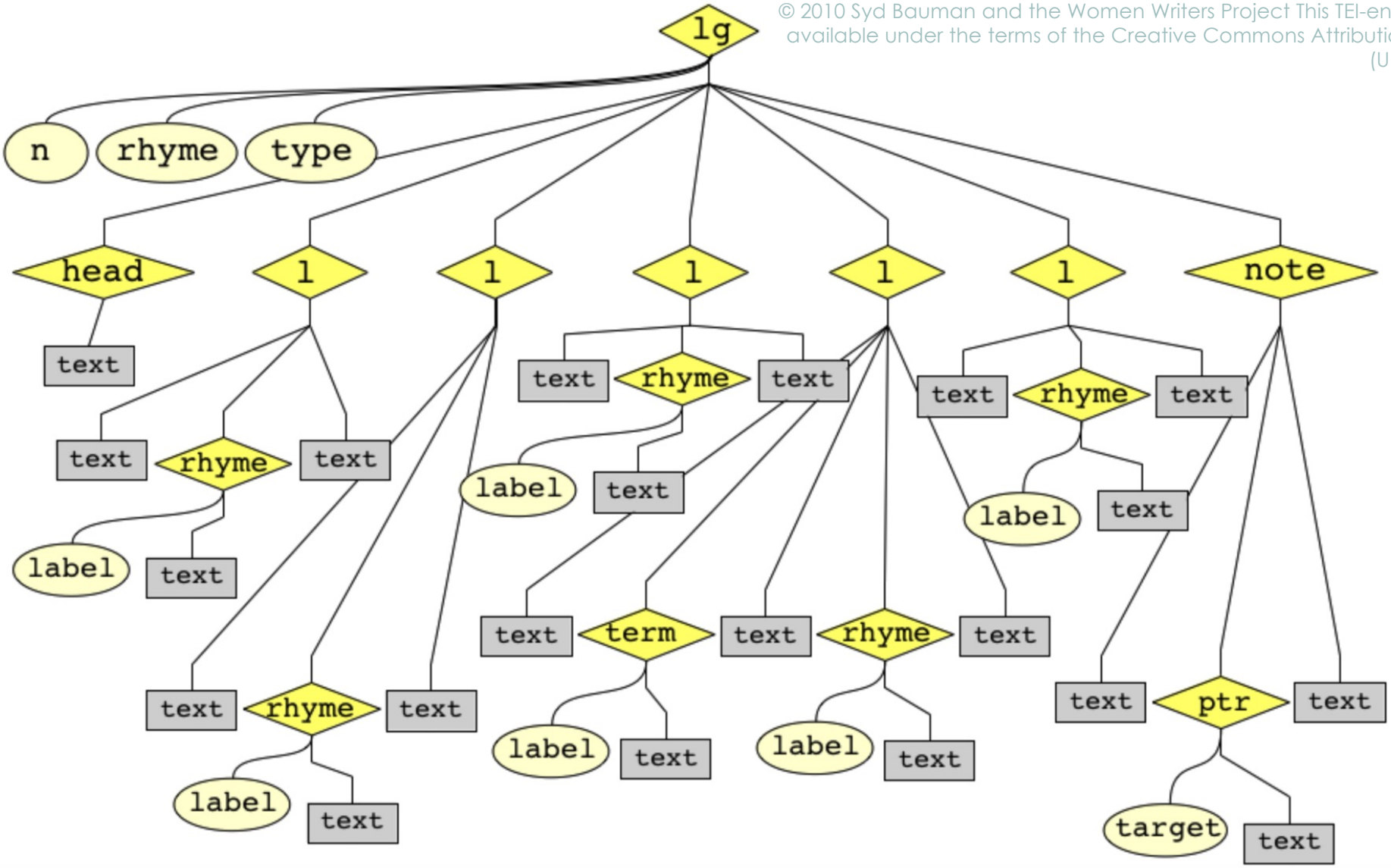
1. geographical coordinates encoded in the authority list, how many did you find?
2. names of people mentioned in the text, how many did you find?
3. names of places mentioned in the text, how many did you find?

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>
```

Simplified XML tree





Don’t want them all?

- Many of the above XPaths return multiple nodes — what if you only want a particular one?
- 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.

Predicates

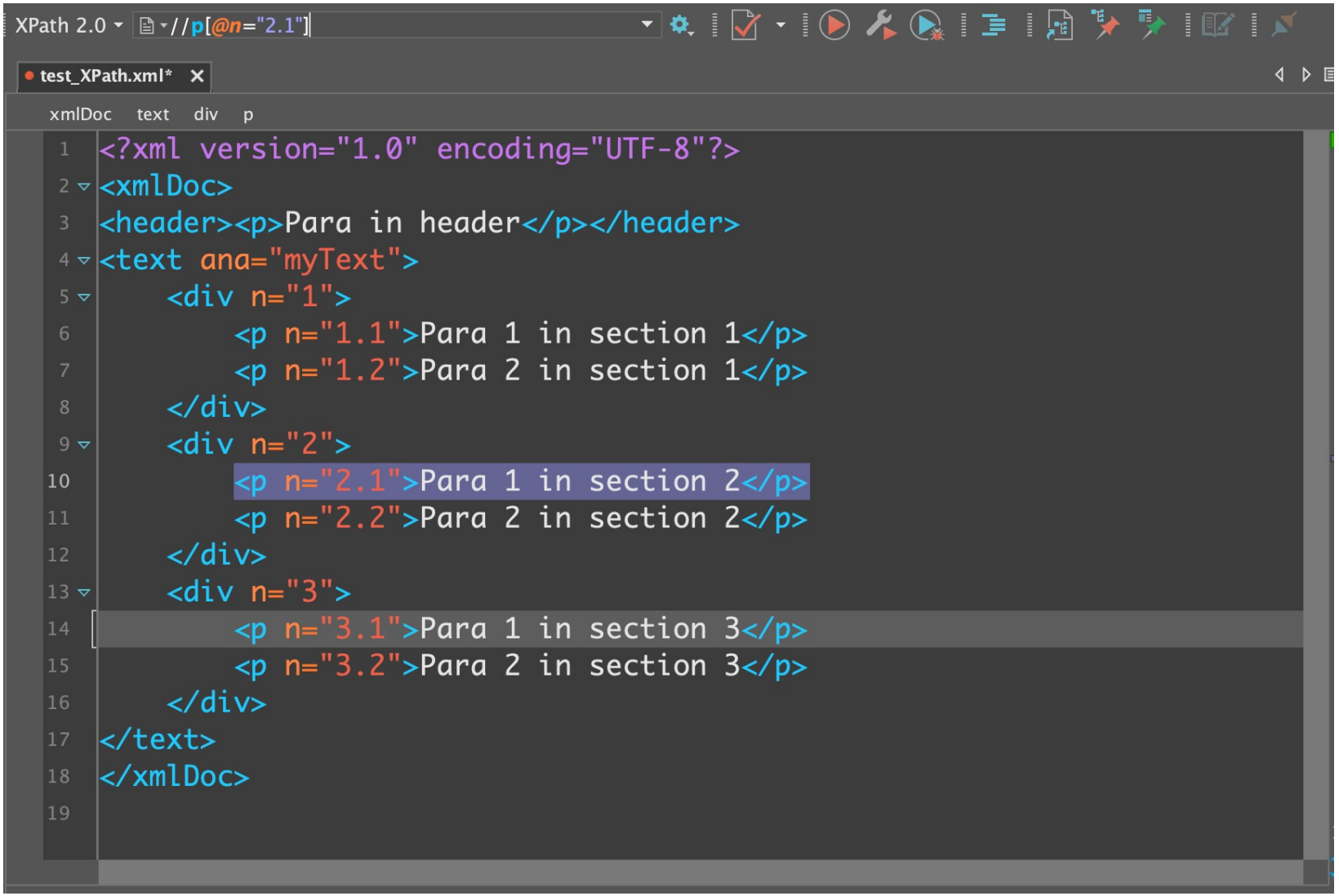
An XPath *predicate* filters the nodes retrieved by a given step

Predicates are expressed after the node test in square brackets

The following are based on http://www.wwp.neu.edu/outreach/seminars/cust_2011-08/demos/xslt_intro/places.xml

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)

//p[@n="2.1"]


XPath 2.0
//p[@n="2.1"]
test_XPath.xml*
xmlDoc text div p
<?xml version="1.0" encoding="UTF-8"?>
<xmlDoc>
<header> Para in header </header>
<text ana="myText">

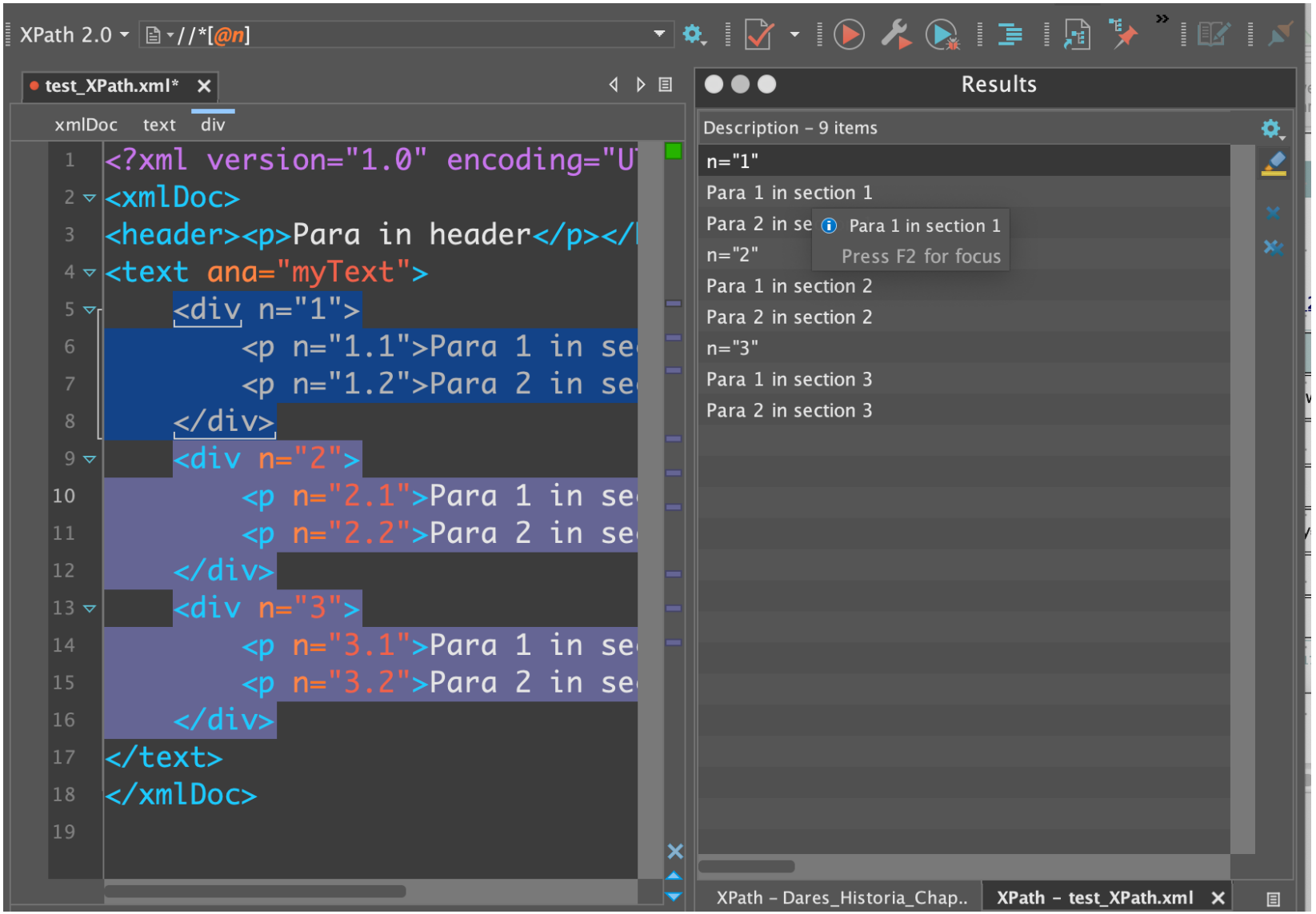
Para 1 in section 1
Para 2 in section 1

Para 1 in section 2
Para 2 in section 2

Para 1 in section 3
Para 2 in section 3

</text>
</xmlDoc>

//*[@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]

XPath 2.0 //text//p[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>
```

//text/descendant::p[1]

XPath 2.0 //text/descendant::p[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>
```

Translate

1. //book[@category='fiction']
2. //*[@type]
3. //chapter[5]/s[1]

Translate

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'.

Exercise 2

- Using Dares_Historia_Chapter_14_annotations_spoiler.xml (Week8/Ex), find out:
  1. which place name is the first one mentioned in the chapter
  2. which place name is the second one mentioned in the chapter
  3. which place name is the first one mentioned in the third sentence
  4. which person name is the fourth mentioned in the chapter.
  5. find out how many ships brought Achilles

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_Bayerische_Staatsbibliothek_Clm_305_transcr_spoiler.xml (Week6/Ex1)
  - **Group 2:** Dares_edition_spoiler.xml (Week7/Ex1)
  - **Group 3:** Ex1_Paris_BnF_Latin_5691_Description_NKY.xml (Week5/Ex1)
  - **Group 4:** Ex2_FirstFolio_spoiler_Bod.xml (Week3/Exercises)

Exercise 3.2

- Post your questions in: https://tinyurl.com/XPathQuiz