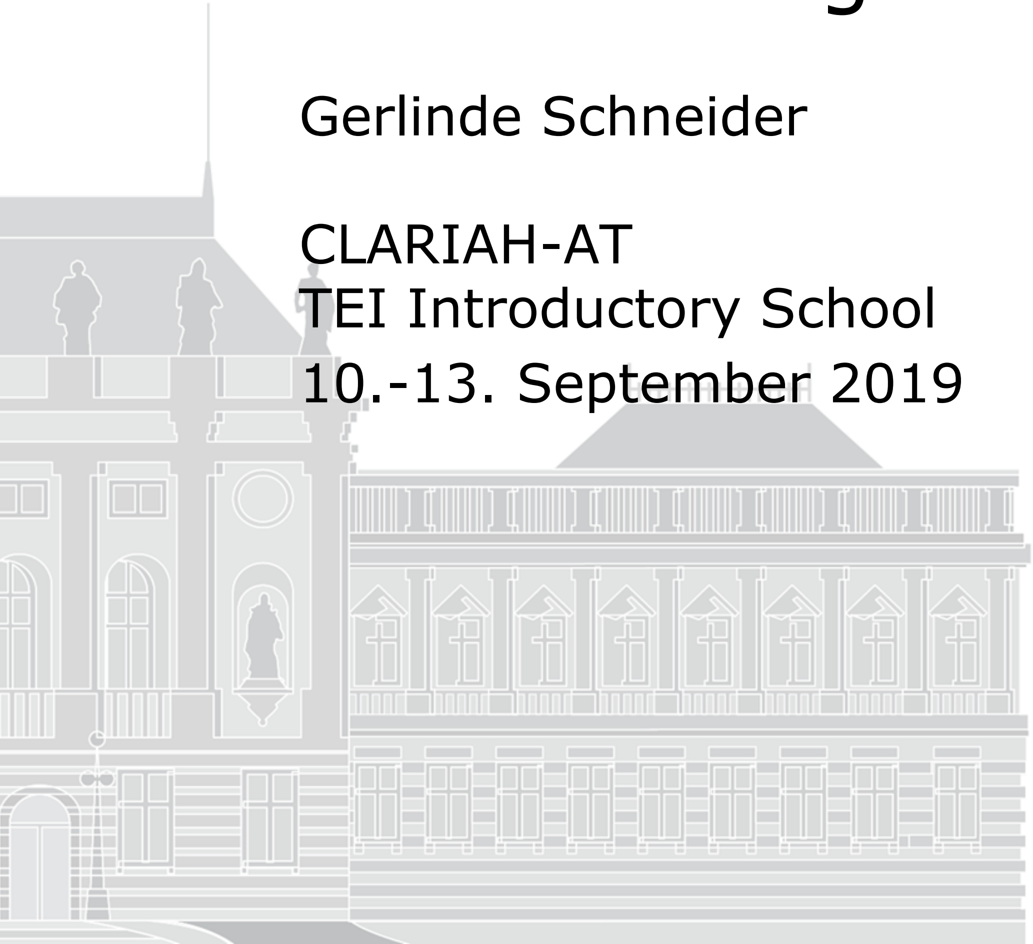
Textencoding with XML

Gerlinde Schneider

CLARIAH-AT
TEI Introductory School
10.-13. September 2019



Venue:
Centre for Information Modelling (ZIM)
Elisabethstraße 59/III, SR 81.31

Programm

Tuesday 10.9.

9:00-10:30	*Digital Scholarly Editing* Roman Bleier, Martina Scholger
11:00-12:30	*Textencoding with XML* Hans Clausen, Gerlinde Schneider
14:00-15:30	*TEI – General Introduction* Hans Clausen, Gerlinde Schneider
16:00-17:30	*Practice: XML/TEI* Hans Clausen, Gerlinde Schneider

Schedule

★ What is XML?

★ Why Text Encoding?

★ XML Basics

★ XML Syntax and Rules

★ Using the Oxygen XML editor

★ A beginners exercise

Goal: To understand

- how XML works and the relationship between XML and TEI

What is XML, actually?

XML stands for e**X**tensible **M**arkup **L**anguage

A generic standard for the **description** and **exchange**
text documents / textual data

W3C Standard

- 1.1 (Second Edition) - currently in use (4th edition of version 1.0) - 2006; 12 years ago

  https://www.w3.org/TR/2006/REC-xml11-20060816/

Why XML

- ❑ System and platform independent
- ❑ Human and machine-readable
- ❑ Low-threshold
- ❑ Supported by a wide range of software
- ❑ International user and developer community
- ❑ Encompasses a whole range of accompanying standards

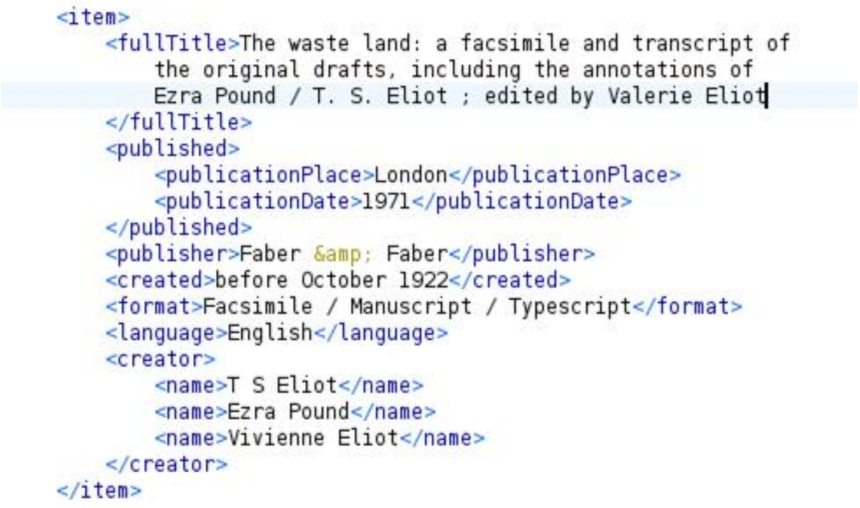
... separates structure and presentation

XML is extensible:
No predefined structure or names of elements and attributes,
Easily adaptable to the needs of specific domains and use cases

Model structured datasets

- Often used in software development and Information Sciences
- Data exchange and storage

E.g. Configuration files, metadata records ...



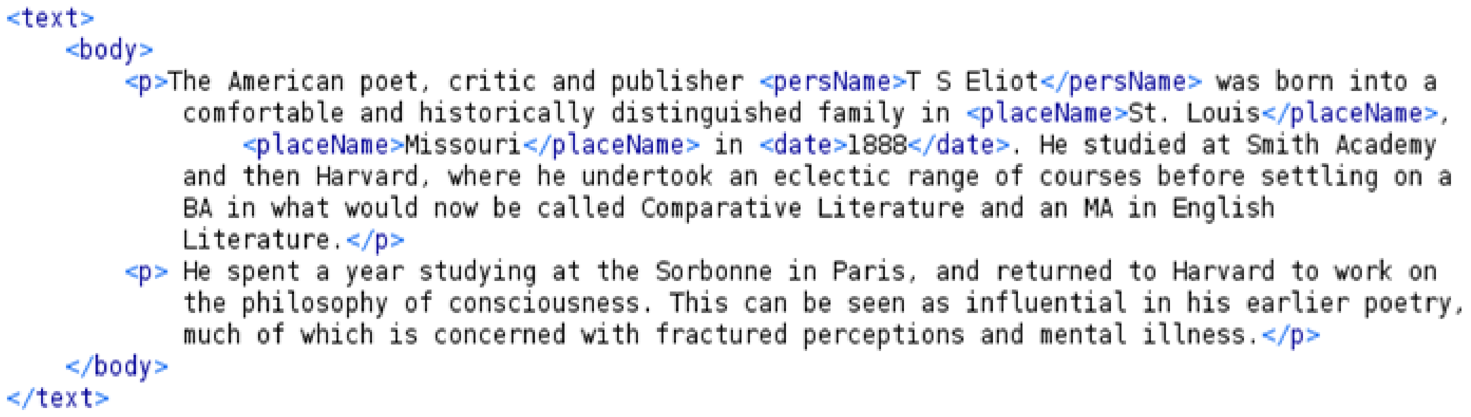
```
<item>
    <fullTitle>The waste land: a facsimile and transcript of
        the original drafts, including the annotations of
        Ezra Pound / T. S. Eliot ; edited by Valerie Eliot
    </fullTitle>
    <published>
        <publicationPlace>London</publicationPlace>
        <publicationDate>1971</publicationDate>
    </published>
    <publisher>Faber &amp; Faber</publisher>
    <created>before October 1922</created>
    <format>Facsimile / Manuscript / Typescript</format>
    <language>English</language>
    <creator>
        <name>T S Eliot</name>
        <name>Ezra Pound</name>
        <name>Vivienne Eliot</name>
    </creator>
</item>
```

Source: British Library

Model narrative text

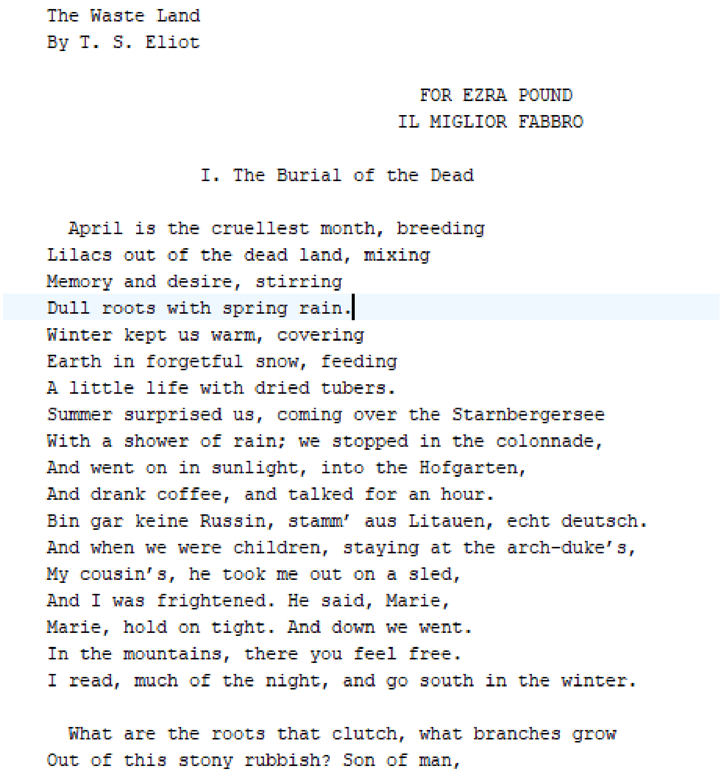
Very important for text-based humanities

Mixed content: Elements can contain strings without markup + other elements



```
<text>
    <body>
        <p>The American poet, critic and publisher <persName>T S Eliot</persName> was born into a
            comfortable and historically distinguished family in <placeName>St. Louis</placeName>,
                <placeName>Missouri</placeName> in <date>1888</date>. He studied at Smith Academy
            and then Harvard, where he undertook an eclectic range of courses before settling on a
            BA in what would now be called Comparative Literature and an MA in English
            Literature.</p>
        <p> He spent a year studying at the Sorbonne in Paris, and returned to Harvard to work on
            the philosophy of consciousness. This can be seen as influential in his earlier poetry,
            much of which is concerned with fractured perceptions and mental illness.</p>
    </body>
</text>
```

Text encoding



```
The Waste Land
By T. S. Eliot


                              FOR EZRA POUND
                             IL MIGLIOR FABBRO

              I. The Burial of the Dead

  April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade,
And went on in sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.
Bin gar keine Russin, stamm' aus Litauen, echt deutsch.
And when we were children, staying at the arch-duke's,
My cousin's, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, there you feel free.
I read, much of the night, and go south in the winter.

  What are the roots that clutch, what branches grow
Out of this stony rubbish? Son of man,
```

There is much more information in a text than can be expressed by character encoding.

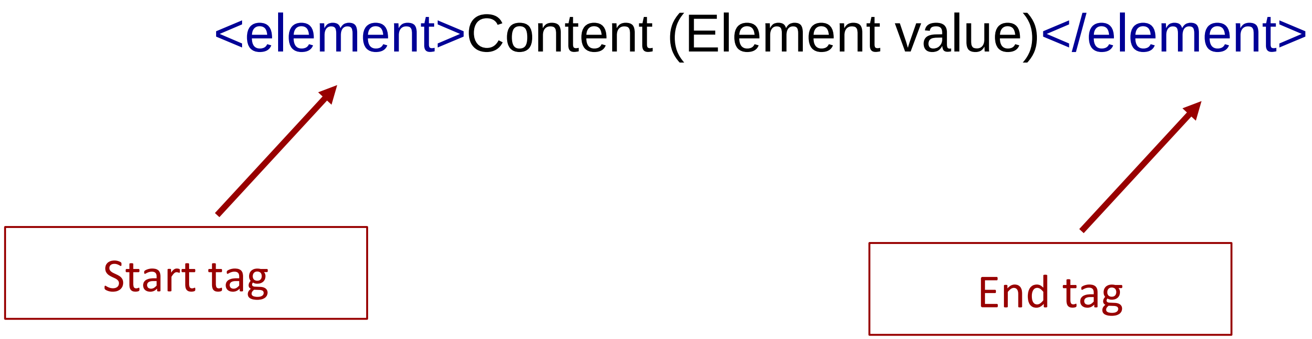
Implicit **content information** or **text structure** are **made explicit (machine-readable)** by markup.

Different **interpretations of a text** and different **readings**, can also be explicated with markup.

Basic syntax

The basic unit is the XML **Element**

- An element is data, surrounded by a tag
- Elements must have a starting and a closing tag
- Elements can contain other elements, text or both



<element>Content (Element value)</element>

Start tag

End tag

Elements

- Empty elements don’t have content and are represented by a special tag

```
<element />
```

Attributes

Attributes give additional information to an element

- Assigned to the start tag of an element
- Name/Value pair
- Elements can take an unlimited number of attributes,
- but only one with one name
- Values must be quoted

```
<element attribute="value">Content</element>
```

Attribute or element

```
<word>
        <text>specific</text>
        <type>adjective</type>
        <lemma>specific</lemma>
</word>
```

OR

```
  <w type="adjective" lemma="specific">specific<w>
```

Rules for XML names

Apply to element and attribute names

- Names can contain any alphanumeric characters, hyphens, dots, or underscores
- Names must begin with an alphabetical character, underscore or colon.
- Names must not start with a number.
- Names are case sensitive: they distinguish between uppercase and lowercase:

  <title> ≠ <Title>

- Names can be of any length.
- The usage of < > & ' and " is not allowed.

Entity references

For reserved characters

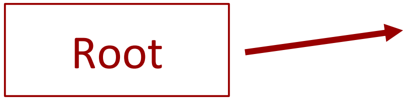
&lt;	<	less than
&gt;	>	greater than
&amp;	&	ampersand
&apos;	'	apostrophe
&quot;	"	quotation mark

Always escape them in your content and don’t use them in XML names

Root Element and Nesting

- An XML document has **one single root element**!

Having elements within another element is called **nesting**



Root

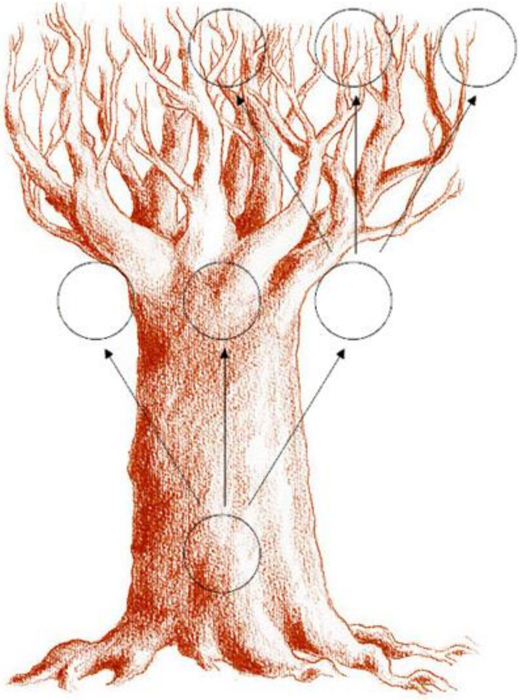
```
<p>A paragraph includes a specific
     <w>word</w> and other words.
     <s>And some sentences.</s>
     <s>And some sentences.</s>
</p>
```

No cross-nesting!

Tree structure

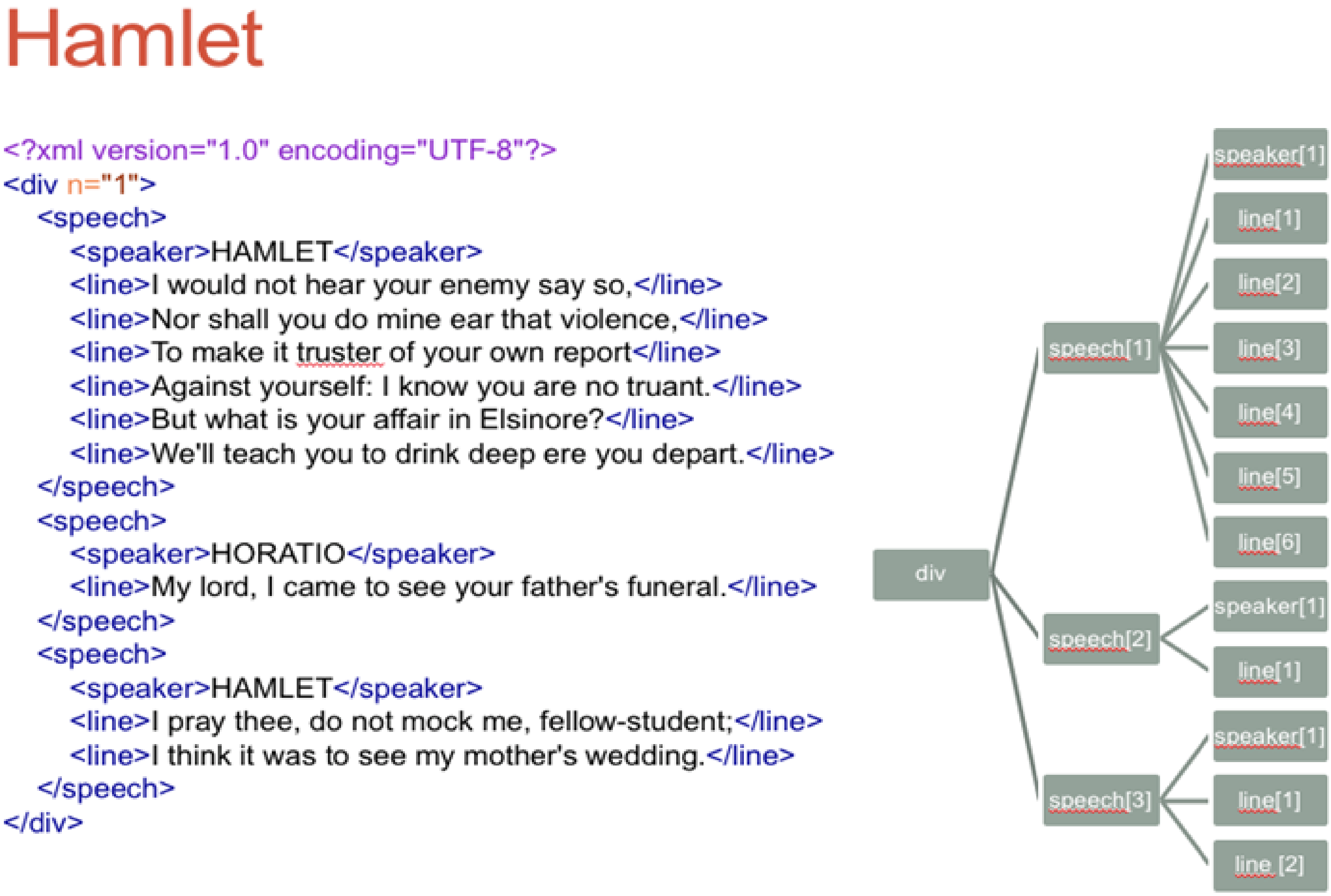
- The nesting of the elements, beginning with one root results in a tree structure

- Ordered Hierarchy of Content Objects (OHCO)
  - Content objects (nodes/elements)
  - Hierarchical (Relation between nodes)
  - ordered (Sequence of nodes)



Baum: ©2008 Ilka Janke

Example



Hamlet

```
<?xml version="1.0" encoding="UTF-8"?>
<div n="1">
   <speech>
      <speaker>HAMLET</speaker>
      <line>I would not hear your enemy say so,</line>
      <line>Nor shall you do mine ear that violence,</line>
      <line>To make it truster of your own report</line>
      <line>Against yourself: I know you are no truant.</line>
      <line>But what is your affair in Elsinore?</line>
      <line>We'll teach you to drink deep ere you depart.</line>
   </speech>
   <speech>
      <speaker>HORATIO</speaker>
      <line>My lord, I came to see your father's funeral.</line>
   </speech>
   <speech>
      <speaker>HAMLET</speaker>
      <line>I pray thee, do not mock me, fellow-student;</line>
      <line>I think it was to see my mother's wedding.</line>
   </speech>
</div>
```

XML Document structure

❑ XML declaration

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

❑ Processing instructions

<?xml-stylesheet type="text/xsl"
href="transformation.xsl"?>

❑ Root element + Nested elements

<root> ... </root>

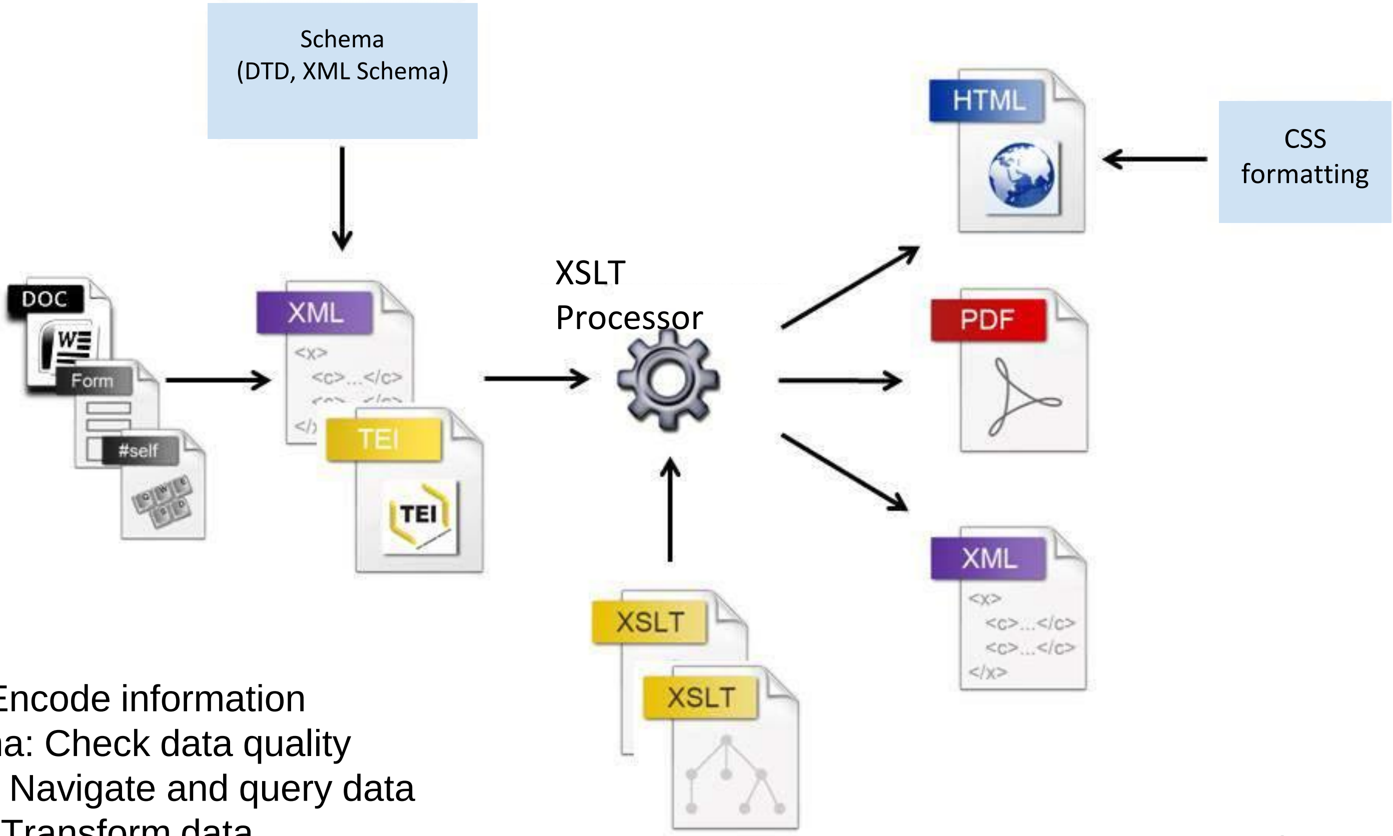
❑ Comments

<!– This is a comment -->

XML Document structure

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="transformation.xsl"?>

<!-- Metadata starts here -->

<item>
        <fullTitle>The waste land: a facsimile and transcript of the original drafts, including the
                        annotations of Ezra Pound / T. S. Eliot ; edited by Valerie Eliot
        </fullTitle>
        <published>
                        <publicationPlace>London</publicationPlace>
                        <publicationDate>1971</publicationDate>
        </published>
        <publisher>Faber &amp; Faber</publisher>
        <created>before October 1922</created>
        <format>Facsimile / Manuscript / Typescript</format>
        <language>English</language>
        <creator>
                        <name>T S Eliot</name>
                        <name>Ezra Pound</name>
                        <name>Vivienne Eliot</name>
                        <!-- Another comment -->
        </creator>
</item>
```

XML Processing



Schema
(DTD, XML Schema)

XSLT
Processor

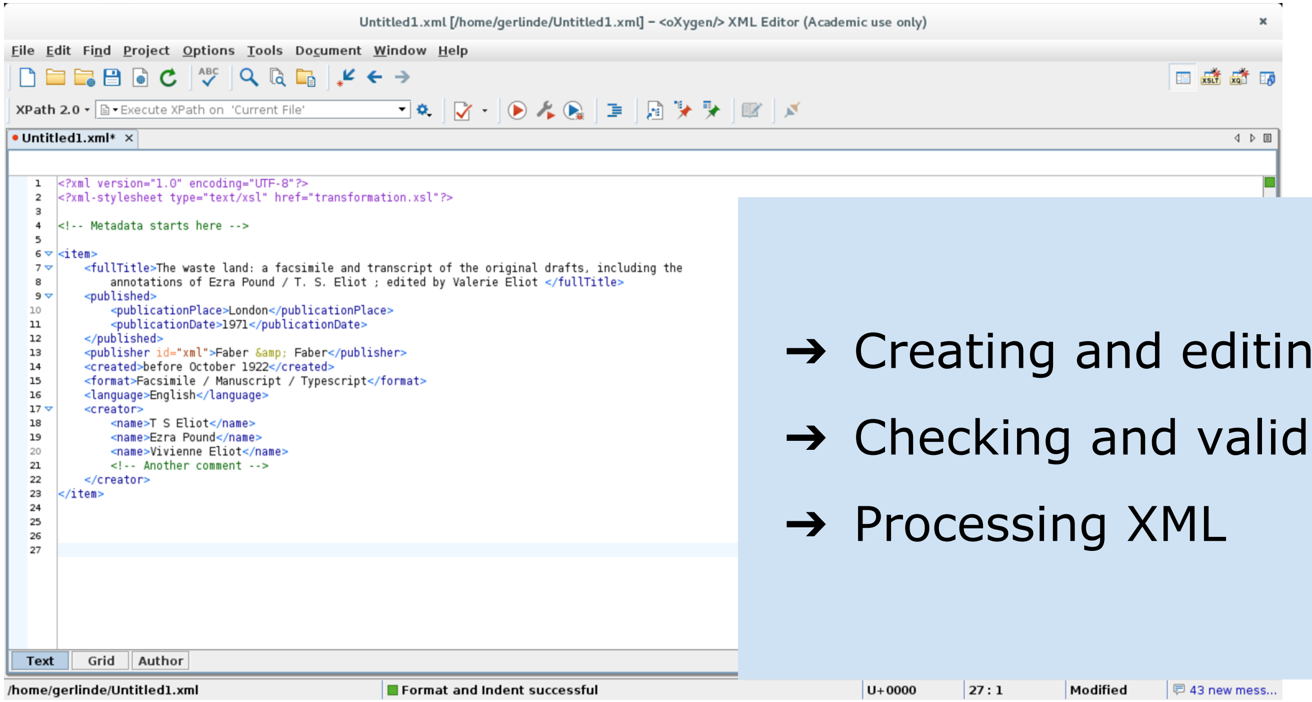
CSS
formatting

XML: Encode information
Schema: Check data quality
XPath: Navigate and query data
XSLT: Transform data

Source: Fritze, 2017

Oxygen editor

Oxygen is a text editor/ development environment specifically for the editing of XML Documents.


Untitled1.xml [/home/gerlinde/Untitled1.xml] - <oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0
Execute XPath on 'Current File'
Untitled1.xml*
Text
Grid
Author

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="transformation.xsl"?>

<!-- Metadata starts here -->

<item>
    <fullTitle>The waste land: a facsimile and transcript of the original drafts, including the
        annotations of Ezra Pound / T. S. Eliot ; edited by Valerie Eliot </fullTitle>
    <published>
        <publicationPlace>London</publicationPlace>
        <publicationDate>1971</publicationDate>
    </published>
    <publisher id="xml">Faber &amp; Faber</publisher>
    <created>before October 1922</created>
    <format>Facsimile / Manuscript / Typescript</format>
    <language>English</language>
    <creator>
        <name>T S Eliot</name>
        <name>Ezra Pound</name>
        <name>Vivienne Eliot</name>
        <!-- Another comment -->
    </creator>
</item>
```

/home/gerlinde/Untitled1.xml | Format and Indent successful | U+0000 | 27 : 1 | Modified | 43 new mess...

- ➔ Creating and editing of XML
- ➔ Checking and validating XML
- ➔ Processing XML

Oxygen editor

- Platform-independent
- Subversion client, Add-on for Git integration
- Supports TEI
- Supports all popular schema languages
- Syntax completion
- Integrated documentation
- Built-in XSLT and FOP processors

Is your document well-formed?

A well-formed XML document fulfills the rules of the standard:

1. There is exactly one root element
2. Each element has a start-tag and an appropriate closing-tag
3. Elements have to be properly nested - no overlapping structures
4. Attribute values have to be quoted
5. An element cannot have two attributes with the same name
6. Reserved characters have to be escaped

Harold/Means, 2004

Is your document well-formed?

- <name>Franz Kafka</name> ➕
- <name><forename>Franz</forename><surname>Kafka</surname></name> ➕
- <name><forename>Franz<surname></forename>Kafka</surname></name> ➖
- <name type="person">Franz Kafka</name> ➕
- <name type=person>Franz Kafka</name> ➖
- <name type="person">Franz Kafka<name/> ➖
- <name type="person">Franz Kafka</Name> ➖
- <name>Franz Kafka<person/></name> ➕
- <name type="person" type="schriftsteller">Franz Kafka</name> ➖
- <name type="person author">Franz Kafka</name> ➕

Validation

In addition to being well-formed a document can also be checked for being **valid**

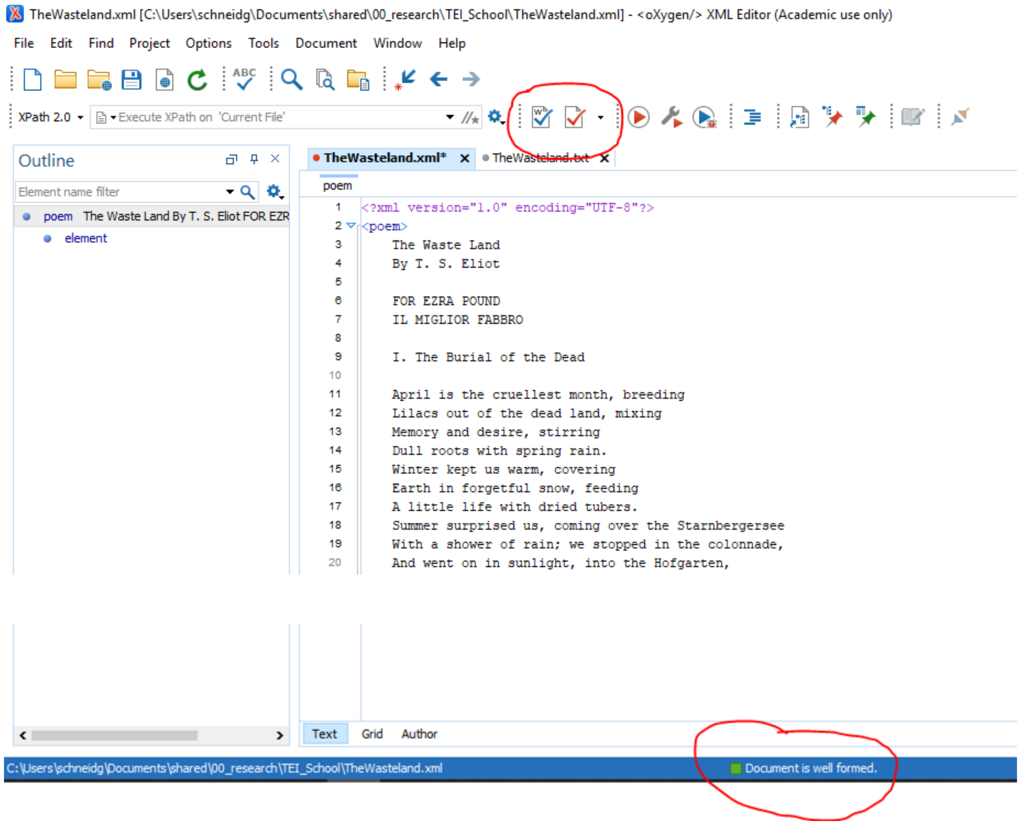
A valid XML document fulfills a set of rules defined in a specific schema, attached to it,

which for example defines...

- the vocabulary (element and attribute names) used
- the structure of a document and the sequence of elements

Different schema languages, e.g. Document Type Definition (DTD) or XMLSchema, allow different types of validation

Document check

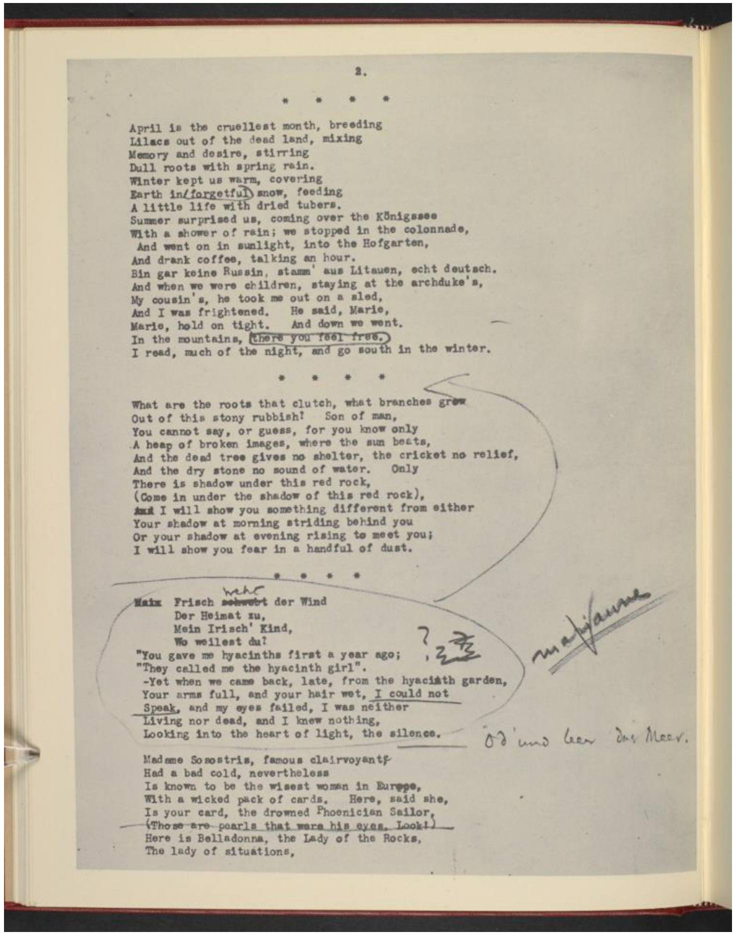

TheWasteland.xml [C:\Users\schneidg\Documents\shared\00_research\TEI_School\TheWasteland.xml] - <oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0
Execute XPath on 'Current File'
Outline
Element name filter
poem The Waste Land By T. S. Eliot FOR EZR
element
TheWasteland.xml*
TheWasteland.txt
poem
Text Grid Author

```
<?xml version="1.0" encoding="UTF-8"?>
<poem>
    The Waste Land
    By T. S. Eliot

    FOR EZRA POUND
    IL MIGLIOR FABBRO

    I. The Burial of the Dead

    April is the cruellest month, breeding
    Lilacs out of the dead land, mixing
    Memory and desire, stirring
    Dull roots with spring rain.
    Winter kept us warm, covering
    Earth in forgetful snow, feeding
    A little life with dried tubers.
    Summer surprised us, coming over the Starnbergersee
    With a shower of rain; we stopped in the colonnade,
    And went on in sunlight, into the Hofgarten,
```

C:\Users\schneidg\Documents\shared\00_research\TEI_School\TheWasteland.xml

🟩 Document is well formed.

Exercise



2.

* * * *

April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in/forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Königssee
With a shower of rain; we stopped in the colonnade,
And went on in sunlight, into the Hofgarten,
And drank coffee, talking an hour.
Bin gar keine Russin, stamm' aus Litauen, echt deutsch.
And when we were children, staying at the archduke's,
My cousin's, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, ~~there you feel free.~~
I read, much of the night, and go south in the winter.

* * * *

What are the roots that clutch, what branches grow
Out of this stony rubbish! Son of man,
You cannot say, or guess, for you know only
A heap of broken images, where the sun beats,
And the dead tree gives no shelter, the cricket no relief,
And the dry stone no sound of water. Only
There is shadow under this red rock,
(Come in under the shadow of this red rock),
~~And~~ I will show you something different from either
Your shadow at morning striding behind you
Or your shadow at evening rising to meet you;
I will show you fear in a handful of dust.

* * * *

~~Main~~ Frisch ~~schwebt~~ weht der Wind
Der Heimat zu,
Mein Irisch' Kind,
Wo weilest du?
"You gave me hyacinths first a year ago; ?
"They called me the hyacinth girl".
-Yet when we came back, late, from the hyacinth garden,
Your arms full, and your hair wet, I could not
Speak, and my eyes failed, I was neither
Living nor dead, and I knew nothing,
Looking into the heart of light, the silence.
Od' und leer das Meer.

Madame Sosostris, famous clairvoyante
Had a bad cold, nevertheless
Is known to be the wisest woman in Europe,
With a wicked pack of cards. Here, said she,
Is your card, the drowned Phoenician Sailor,
~~(Those are pearls that were his eyes. Look!)~~
Here is Belladonna, the Lady of the Rocks,
The lady of situations,

The Wasteland
by T. S. Eliot

There are many interesting entities and phenomena to markup in this poem!

Source: British Library

Exercise

- Start the Oxygen XML editor
- Open a new XML Document

  Strg+N | File -> New File | 🗋 -> XML Document

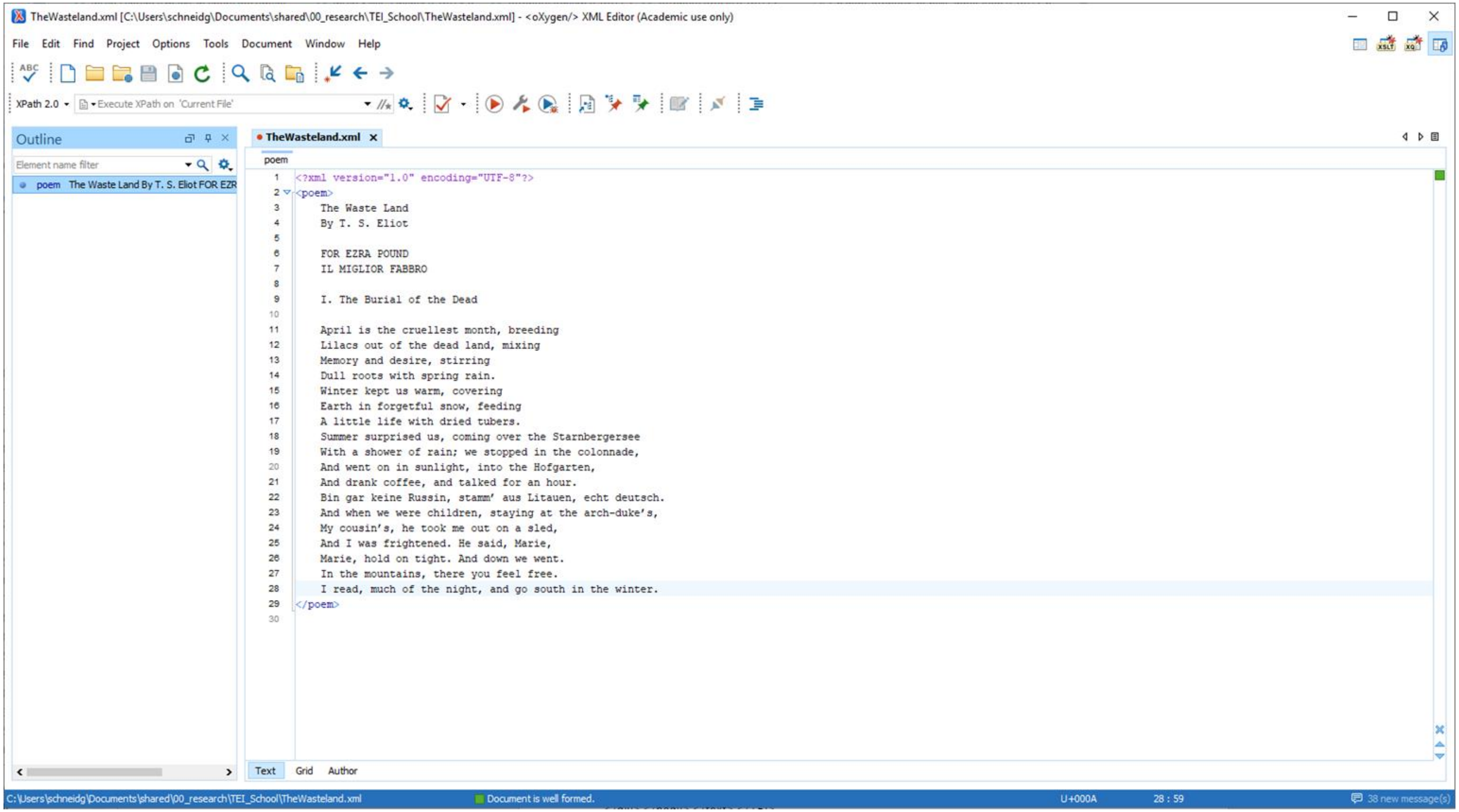
- Create a root element
- Copy the text from the File ‘TheWasteland.txt’ into your root element
- Describe the structural elements and layout of the first part of the poem *(I. The Burial of the Dead)* with meaningful XML elements and attributes

Brainstorming:

Which relevant structural and content-related elements of the text can you identify?

Which metadata could be added to the text?

Exercise


TheWasteland.xml [C:\Users\schneidg\Documents\shared\00_research\TEI_School\TheWasteland.xml] - <oXygen/> XML Editor (Academic use only)
File
Edit
Find
Project
Options
Tools
Document
Window
Help
XPath 2.0
Execute XPath on 'Current File'
Outline
Element name filter
poem The Waste Land By T. S. Eliot FOR EZR
TheWasteland.xml
poem
Text
Grid
Author
C:\Users\schneidg\Documents\shared\00_research\TEI_School\TheWasteland.xml
U+000A
28 : 59
38 new message(s)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<poem>
    The Waste Land
    By T. S. Eliot

    FOR EZRA POUND
    IL MIGLIOR FABBRO

    I. The Burial of the Dead

    April is the cruellest month, breeding
    Lilacs out of the dead land, mixing
    Memory and desire, stirring
    Dull roots with spring rain.
    Winter kept us warm, covering
    Earth in forgetful snow, feeding
    A little life with dried tubers.
    Summer surprised us, coming over the Starnbergersee
    With a shower of rain; we stopped in the colonnade,
    And went on in sunlight, into the Hofgarten,
    And drank coffee, and talked for an hour.
    Bin gar keine Russin, stamm’ aus Litauen, echt deutsch.
    And when we were children, staying at the arch-duke’s,
    My cousin’s, he took me out on a sled,
    And I was frightened. He said, Marie,
    Marie, hold on tight. And down we went.
    In the mountains, there you feel free.
    I read, much of the night, and go south in the winter.
</poem>
```

Document is well formed.

Material

- Manuscript Facsimile at the British Library:
  https://www.bl.uk/collection-items/manuscript-of-t-s-eliots-the-waste-land-with-ezra-pounds-annotations
- Full text at the Poetry Foundation:
  https://www.poetryfoundation.org/poems/47311/the-waste-land

Finished?

- Check if your XML document is well-formed.

  Ctrl + Shift + w | ✓

- Save your work on your computer

  Ctrl + s | 💾

- Upload your files to the ‘Exercises’ folder so that we can compare them - https://tinyurl.com/Textencoding-with-XML 😁

  - What did you mark up?
  - Did you use elements or attributes?

References

XML in a Nutshell by Elliotte Rusty Harold, W. Scott Means 2004 by O'Reilly Media, Third edition

Textkodierung mit XML, Summer School “Digitale Edition” 2017, Christiane Fritze

Manuscript of T S Eliot's The Waste Land, with Ezra Pound's annotations, British Library, https://www.bl.uk/collection-items/manuscript-of-t-s-eliots-the-waste-land-with-ezra-pounds-annotations

https://www.poetryfoundation.org/poems/47311/the-waste-land

http://www.w3schools.com/xml/