

**TALOS ERA CHAIR IN ARTIFICIAL INTELLIGENCE  
FOR HUMANITIES AND SOCIAL SCIENCES**



ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ  
UNIVERSITY OF CRETE



**LEAF**

Linked Editing  
Academic Framework

An open-source tool for text  
encoding

Rachel Milio

PhD Candidate, RT1  
TALOS-AI4SSH

rachel.milio@outlook.com



Funded by the  
European Union



Horizon ERA Chair TALOS AI4SSH Project funded by the European Commission  
Grant Agreement n° 101087269, <https://cordis.europa.eu/project/id/101087269>

TALOS ERA Chair AI for SSH – Project n° 101087269

"LEAF-Writer", Rachel Milio

CC BY-NC-ND



Hello everyone, and welcome to this AI4SSH MOOC session where we will introduce the Linked Editing Academic Framework, an open-source web-based tool for text encoding.



## What is *text encoding*?

Making human  
readable text  
*machine readable*

AKA  
“markup”

### Types of mark-up:

- Structural
- Presentational
- Semantic



First, it is necessary to introduce the field of text encoding. Through text encoding, we can make human-readable texts machine-readable. Another term for text encoding is *mark-up*. There are three main types of text mark-up: structural, presentational, and semantic. Structural markup denotes the structure of a text (such as headings and paragraphs), presentational markup conveys stylistic choices such as color and font, and semantic markup conveys meaning (such as the name of people and places within a text).



# What is XML?

**XML = eXtensible Markup Language**

- Structure and label data and metadata
- Extensible = no predefined tags

## Extensible Markup

Pros	Cons
Flexibility and freedom for users	Less interoperability/reusability

```
<place>Crete</place>  
<loc>Crete</loc>  
<placename>Crete</placename>
```



The markup language XML can be used for both structural and semantic markup. XML stands for the eXtensible Markup Language, and is used to structure and label data and metadata in texts. However, the extensible part of XML means that it has no predefined tags, giving users freedom and flexibility but also impacting the interoperability and reusability of XML-encoded data.



## What is TEI?

Text Encoding Initiative

Humanistic  
standard for  
XML

Used for literary & linguistic texts  
Manuscripts, historical archives,  
critical editions, anthologies,  
inscriptions...

## Why?

More consistent text  
encoding  
Increase interoperability and  
reusability



TEI, or the Text Encoding Initiative, is a humanistic standard for XML text encoding. Through the use of TEI, text encoding can be more consistent, increasing interoperability and reusability. TEI is most commonly used for the digital encoding of literary and linguistic texts, such as manuscripts, historical archives, and critical editions.

The diagram shows the TEI markup for the text "Helen of Sparta". The markup is as follows:

```

<persName
  ref='https://www.wikidata.org/wiki/Q164061'>
<forename>Helen</forename>of
<placeName
  ref='https://www.wikidata.org/wiki/Q5690'>Sparta
</placeName>
</persName>

```

Annotations in the diagram include:

- “Helen of Sparta”**: A box pointing to the entire markup block.
- Original text string**: A box pointing to the text "Helen of Sparta" within the markup.
- persName = full name of person**: A box pointing to the opening <persName> tag.
- forename = first name (“Helen”)**: A box pointing to the <forename>Helen</forename> segment.
- placeName = location (“Sparta”)**: A box pointing to the <placeName>Sparta</placeName> segment.
- ref = attribute for external reference**: A box pointing to the ref attribute in both the persName and placeName tags.
- Attribute = describes an element**: A box explaining the role of an attribute.
- Element = unit of text = tag**: A box explaining the role of an element.

Additional information at the bottom of the slide includes the URL <https://tei-c.org/>, the project name "TALOS ERA Chair AI for SSH – Project n° 101087269", the author "“LEAF-Writer”, Rachel Millio", the license "CC BY-NC-ND", and the slide number "5".

In TEI, an element is a unit of text. These are also known as tags. An attribute describes an element. For example, the sample text “Helen of Sparta” can be tagged as follows:

```

<persName ref='https://www.wikidata.org/wiki/Q164061'>
<forename>Helen</forename>of <placeName
ref='https://www.wikidata.org/wiki/Q5690'>Sparta
</placeName>
</persName>

```

In this case, the nested tags designate the entire string “Helen of Sparta” as the name of a person, while “Helen” is a forename and “Sparta” is a place name. The attribute “ref” is used to point to an external URI, in this case from the authority base Wikidata. For more on TEI, I recommend checking out the TEI guidelines, which are linked on the slide and contain explanations and examples of the available TEI elements and attributes.



## LEAF-Writer

- Web-based text editor
- Free; no configuration or installation needed!
- Supports sharing and collaboration
- TEI schema support
- On-the-fly validation
- Entity tagging



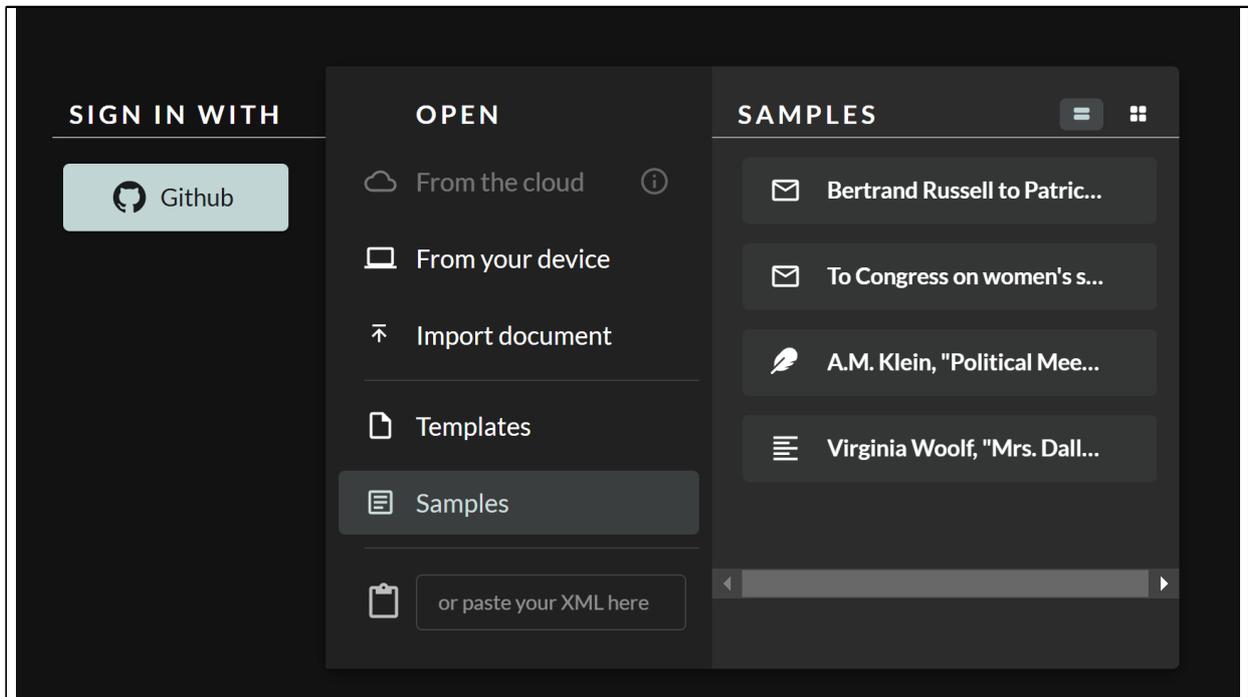
# LEAF

Linked Editing  
Academic Framework

<https://leaf-writer.leaf-vre.org/>



Now that we understand what TEI is and how it is used in the digital humanities, we can introduce the platform LEAF-Writer. LEAF-Writer is a free, web-based semantic text editor. Unlike other XML editors, many of which are proprietary, LEAF-Writer is completely open-source and requires no configuration or installation to use. You can access LEAF-Writer Commons at the link on the screen.



On the LEAF-Writer landing page, you can see the options to open a document. By logging in to Github, you can access documents stored via cloud, but you can also open documents stored locally on your computer or import from the OCR site Transkribus. LEAF-Writer also provides templates and sample documents to start with. Today, we'll work in a sample document, the letter from Bertrand Russell to Patricia Spence. Double click to open the letter.

The screenshot displays the LEAF-Writer web application interface. At the top, the title bar shows 'LEAF WRITER' and the document name 'Bertrand Russell to Patricia Spence (letter).xml'. Below this is a navigation bar with tabs for 'Table of Contents', 'Markup', and 'Entities'. The left sidebar shows a tree view of the document structure, including 'teiHeader', 'text', 'div', 'head', 'opener', 'p', 'pb', and 'closer'. The main editing area shows the document content with TEI-XML tags. The right sidebar contains tabs for 'Raw XML', 'Image Viewer', and 'Validation', with the 'Raw XML' tab active, showing the XML code for the document.

This is the editing view of LEAF-Writer. On the left-hand side, we have three tabs: the Table of Contents, the Markup, and the Entities. The “table of contents” displays section headers (for example, chapter titles in a longer text). “Markup” displays the TEI-XML elements in their nested structure. “Entities” displays the entities which have been tagged and disambiguated. In the right side menu, you can access the Raw XML, allowing for more fine grain editing. You can also view associated images (for example, if a document is a digitized version of a manuscript). Lastly, you can validate your document, ensuring that your encoding complies with the TEI guidelines. In the center top menu, there are the options for types of entities you can tag. For example, this icon corresponds to a person entity.

LEAF WRITER Bertrand Russell to Patricia Spence (letter).xml editing

Table of Contents Markup Entities Tag

TEI

- tailHeader
- text
- body
  - div
    - head
    - opener
    - p I Have Had No Letter Fro
    - p
    - pb
    - p
    - p
    - closer

**Bertrand Russell to Patricia Spence - October 21, 1935**

NOTE:  
Bad writing due to shaky train

In train  
Oslo to Bergen

21.10.35

Dearest -

I have had no letter from you since I left **Stockholm**, but I had a nice one from John in an envelope you had sent him. I had sent him one addressed to Copenhagen but he hadn't used it.

When I reached Oslo yesterday evening, Brynjulf Bull should have been there to meet me, but wasn't. He is not on the telephone, so I took a taxi to his address, which turned out to be a students' club with no one about on Sundays, so I went to a hotel feeling rather non-plussed. But presently he turned up. He had got the

time of my arrival wrong, and when [sic] when he had found he had missed me he phoned to every hotel in Oslo till he hit on the right one. He left me at 10, and then I had to do a Sunday Referee article. Today my journey lasts from 9 till 9 - fortunately one of the most beautiful railway journeys in the world. Tomorrow I lecture at Bergen to the Anglo-Norwegian Society. Next day I go back to Oslo, lecture there Fri. and Sat. and then start for home via Bergen.

MARKUP & LINKING JSON-LD TEI ALL

Bugs / Requests LEAF-Writer 3.6.0 Powered by Tiny

TALOS ERA Chair AI for SSH – Project n° 101087269 “LEAF-Writer”, Rachel Milio CC BY-NC-ND 9

Let's try out tagging an entity. If we highlight the text "Stockholm" and select the Place icon, you'll be prompted by a pop-up to select the correct place identifier.

The image shows a software interface for tagging a place. On the left, a 'Find Place' window displays search results for 'Stockholm'. The results are categorized under 'WIKIDATA' and include:

- Stockholm (capital and largest city of Sweden) - Wikidata 10
- Stockholm (municipality in Stockholm County, Sweden) - LINC5 2
- Stockholm University (state university of Stockholm, Sweden)
- Stockholm (family name)
- Stockholm County (county in Sweden)
- Stockholm City (urban municipality in Sweden until the end of 1970)
- 1912 Summer Olympics (Games of the V Olympiad, celebrated in Stockholm (Sweden) in 1912)

At the bottom of the search window are buttons for 'CANCEL', 'TAG WITHOUT LINKING', and 'SELECT'. On the right, the 'Tag Place' configuration panel is shown. It includes:

- SELECTED TEXT:** Stockholm
- TAG AS:** Stockholm (source: wikidata)
- LEVEL OF CERTAINTY:** High (selected), Medium, Low, Unknown
- PRECISION OF LOCATION OF PLACE NAME:** High (selected), Medium, Low, Unknown
- CERT Certainty:** high (selected in a dropdown menu)
- KEY:** Stockholm
- Attributes:** A list of attributes including 'ana', 'Analysis', 'cert' (highlighted), 'Certainty', 'change', 'copyOf', 'corresp', 'Corresponds', 'datingMethod', 'datingPoint', 'evidence', 'exclude', 'fac', 'Facsimile', 'from', 'from-custom', 'from-iso', and 'full'.

Buttons for 'CANCEL' and 'OK' are located at the bottom of the configuration panel. At the bottom of the entire screenshot, there is a footer with the text: 'TALOS ERA Chair AI for SSH – Project n° 101087269', '“LEAF-Writer”, Rachel Milio', 'CC BY-NC-ND', and a Creative Commons license icon.

The text “Stockholm” refers to the Swedish capital city Stockholm, so we’ll select that option. Now you can select attributes such as certainty and precision.

The screenshot displays the LEAF-WRITER web application interface. At the top, the document title is "Bertrand Russell to Patricia Spence (letter).xml". The left sidebar contains an "Entities" panel with a search filter set to "All" and a "Sequential" view. A single entity, "Stockholm", is listed with the following details: Standard: Stockholm, URI: <http://www.wikidata.org/entity/Q1754>, cert: high, and precision: high. The main editing area shows the document's content, including a title "Bertrand Russell to Patricia Spence - October 21, 1935", a "NOTE" section, and the main body of the letter. A portion of the letter text is highlighted in orange, corresponding to the "Stockholm" entity in the sidebar. The right sidebar displays the "Raw XML" code, showing the TEI markup for the document, including the title, date, and the highlighted entity. The bottom status bar includes the text "MARKUP & LINKING JSON-LD TEI ALL", "Bugs / Requests LEAF-Writer 3.6.0 Powered by Tiny", "TALOS ERA Chair AI for SSH - Project n° 101087269", "LEAF-Writer", Rachel Milio, and a Creative Commons BY-NC-ND license logo.

Choose OK, and see your entity appear in the text and in the entities panel.

The screenshot shows the LEAF-Writer web application interface. The main document area displays the following text:

**Bertrand Russell to Patricia Spence - October 21, 1935**

NOTE:  
Bad writing due to shaky train  
In train  
Oslo to Bergen

21.10.35

Dearest -

I have had no letter from you since I left **Stockholm**, but I had a nice one from John in an envelope you had sent him. I had sent him one addressed to Copenhagen but he hadn't used it.

When I reached Oslo yesterday evening, Brynjulf Bull should have been there to meet me, but wasn't. He is not on the telephone, so I took a taxi to his address, which turned out to be a students' club with no one about on Sundays, so I went to a hotel feeling rather non-plussed. But presently he turned up. He had got the

time of my arrival wrong, and when [sic] when he had found he had missed me he phoned to every hotel in Oslo till he hit on the right one. He left me at 10, and then I had to do a Sunday Referee article. Today my journey lasts from 9 till 9 - fortunately one of the most beautiful railway journeys in the world. Tomorrow I lecture at Bergen to the Anglo-Norwegian Society. Next day I go back to Oslo, lecture there Fri. and Sat. and then start for home via Bergen.

The interface includes a sidebar on the left with a 'Stockholm' location and a person icon. The top navigation bar shows 'LEAF-WRITER' and the document title. The bottom status bar indicates 'MARKUP & LINKING', 'JSON-LD', 'TEI ALL', and 'CC BY-NC-ND' license.

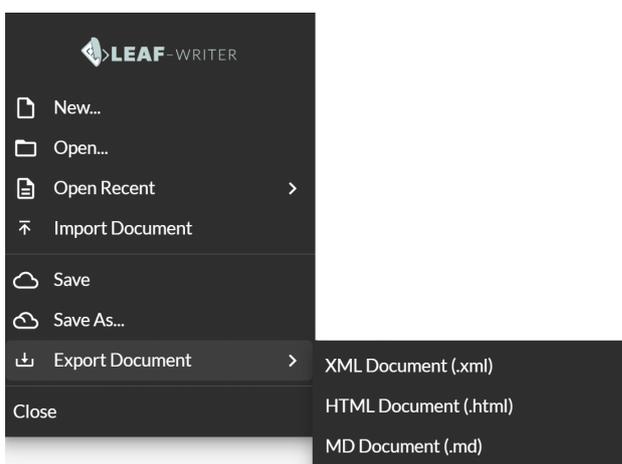
Next, highlight the name Bertrand Russell in the title. Select Person, and choose the appropriate identity.

The screenshot displays the LEAF-Writer web application. The main document area shows a letter from Bertrand Russell to Patricia Spence, dated October 21, 1935. The text includes a note about shaky handwriting, the location (Oslo to Bergen), and a salutation 'Dearest -'. The main body of the letter describes the author's journey from Stockholm to Oslo and then to Bergen. The name 'Bertrand Russell' is highlighted in blue in the first sentence of the main text. The left sidebar shows a list of entities, with 'Bertrand Russell' selected, displaying its standard name, URI, and other metadata. The right sidebar shows the raw XML code for the document, with the corresponding XML tags for the highlighted name visible.

Now, you've tagged the string "Bertrand Russell" as a person entity.



## Exporting from LEAF-Writer



XML: eXtensible  
Markup Language  
HTML: Hypertext  
Markup Language  
MD: Markdown

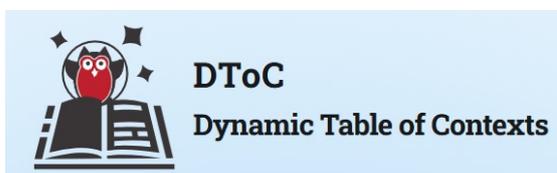
Once you've encoded your text, you have a few options. You can choose to save it to your Github account, therefore storing it in the cloud to return to later. LEAF-Writer autosaves, but you are also able to manually save or save as to change the file name. Additionally, you are able to export the file as XML, HTML, or Markdown.



## To Learn More

<https://www.leaf-vre.org/>

<https://tei-c.org/>



There is so much more to explore with LEAF-Writer and TEI text encoding. LEAF-Writer is only one of multiple tools in the LEAF Commons Suite, including the Dynamic Table of Contexts, an interactive e-reader that combines traditional indices with semantic markup, and NERVE, the Named Entity Reconciliation Vetting Environment, which allows for NER-powered semantic annotation of texts. For more on all that LEAF offers, you can visit the documentation site. There are also plenty of resources for learning more about the TEI guidelines, which are included in the Materials for this course.

**TALOS ERA CHAIR IN ARTIFICIAL INTELLIGENCE  
FOR HUMANITIES AND SOCIAL SCIENCES**



ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ  
UNIVERSITY OF CRETE



Thank you!

Rachel Milio  
PhD Candidate, RT1  
TALOS-AI4SSH  
rachel.milio@outlook.com



Funded by the  
European Union

Horizon ERA Chair TALOS AI4SSH Project funded by the European Commission  
Grant Agreement n° 101087269, <https://cordis.europa.eu/project/id/101087269>

TALOS ERA Chair AI for SSH – Project n° 101087269

"LEAF-Writer", Rachel Milio

CC BY-NC-ND

Thank you so much for joining me to learn about LEAF-Writer for humanistic text encoding.