TALOS ERA CHAIR IN ARTIFICIAL INTELLIGENCE FOR HUMANITIES AND SOCIAL SCIENCES

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

LEAF
Linked Editing Academic Framework

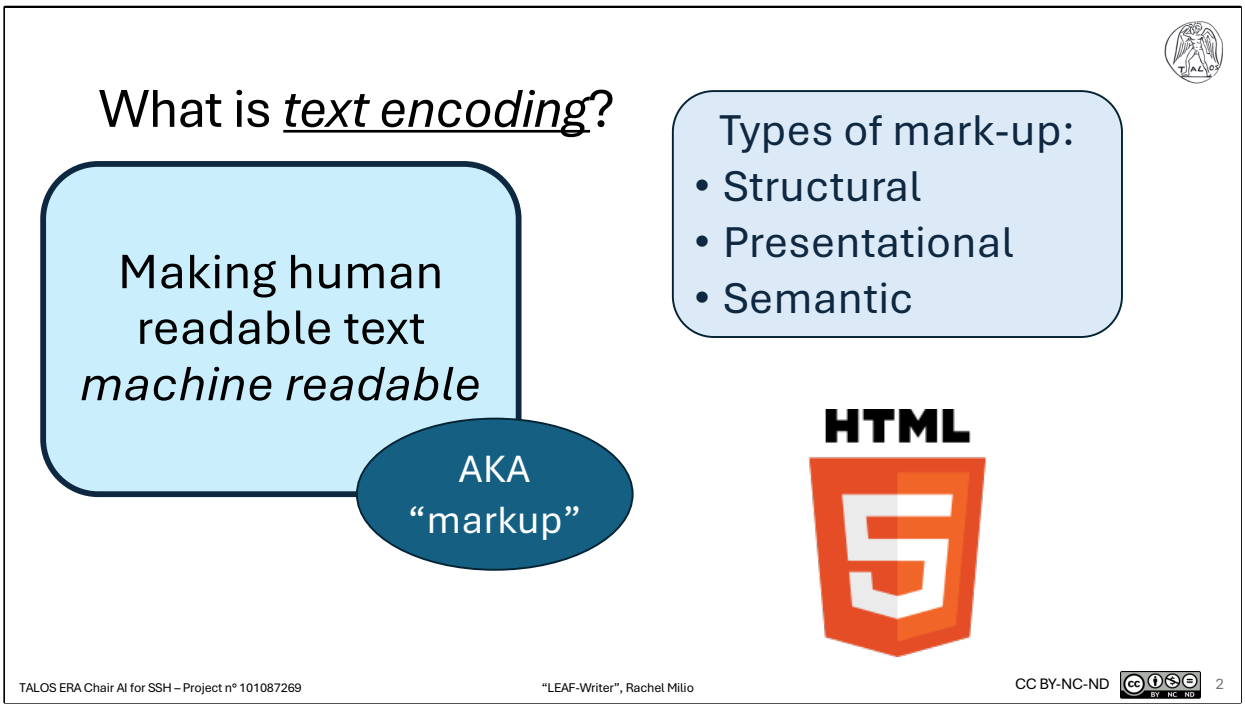
An open-source tool for text encoding

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

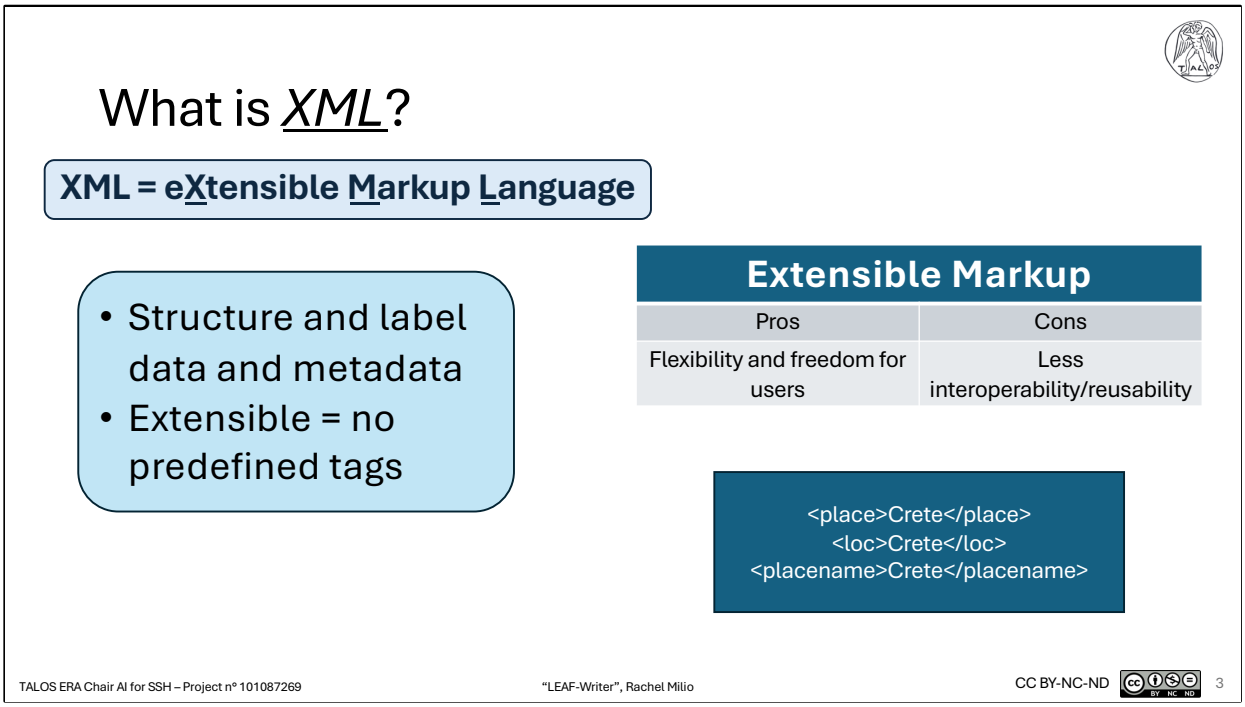
Funded by the European Union

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

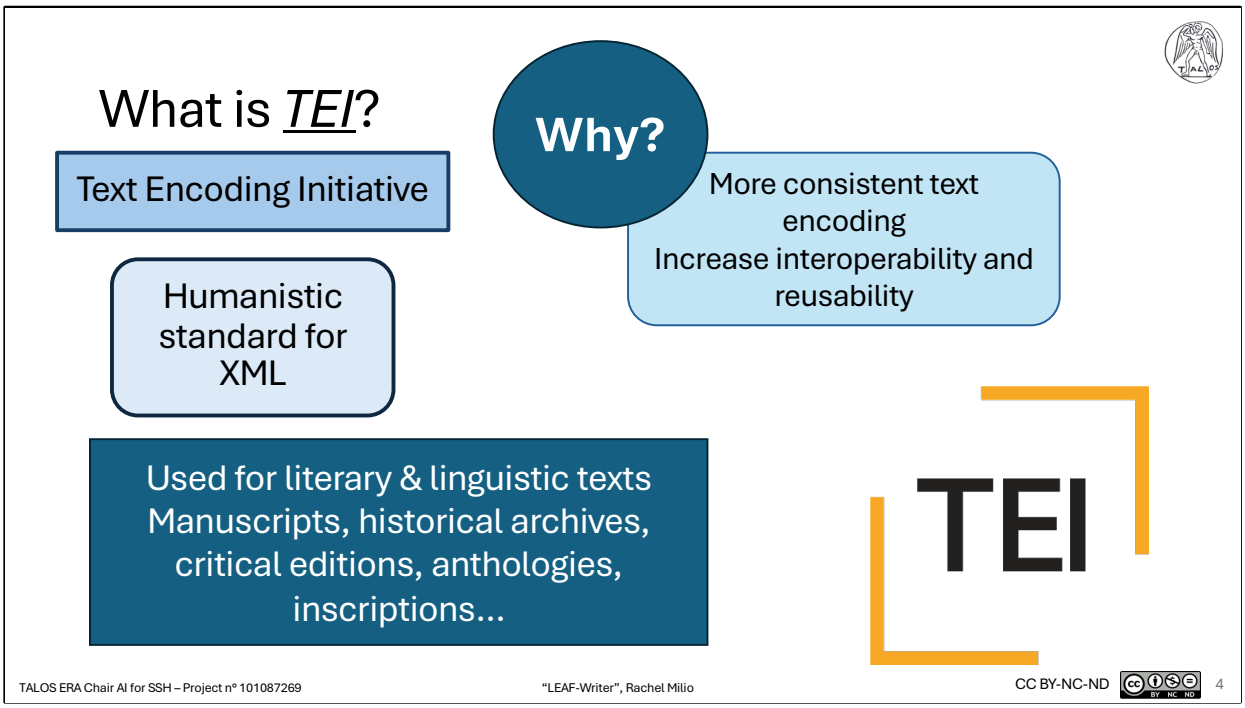
Hello everyone, and welcome to this AI4SSH MOOC session where we will introduce the Linked Editing Academic Framework, an open-source web-based tool for text encoding.



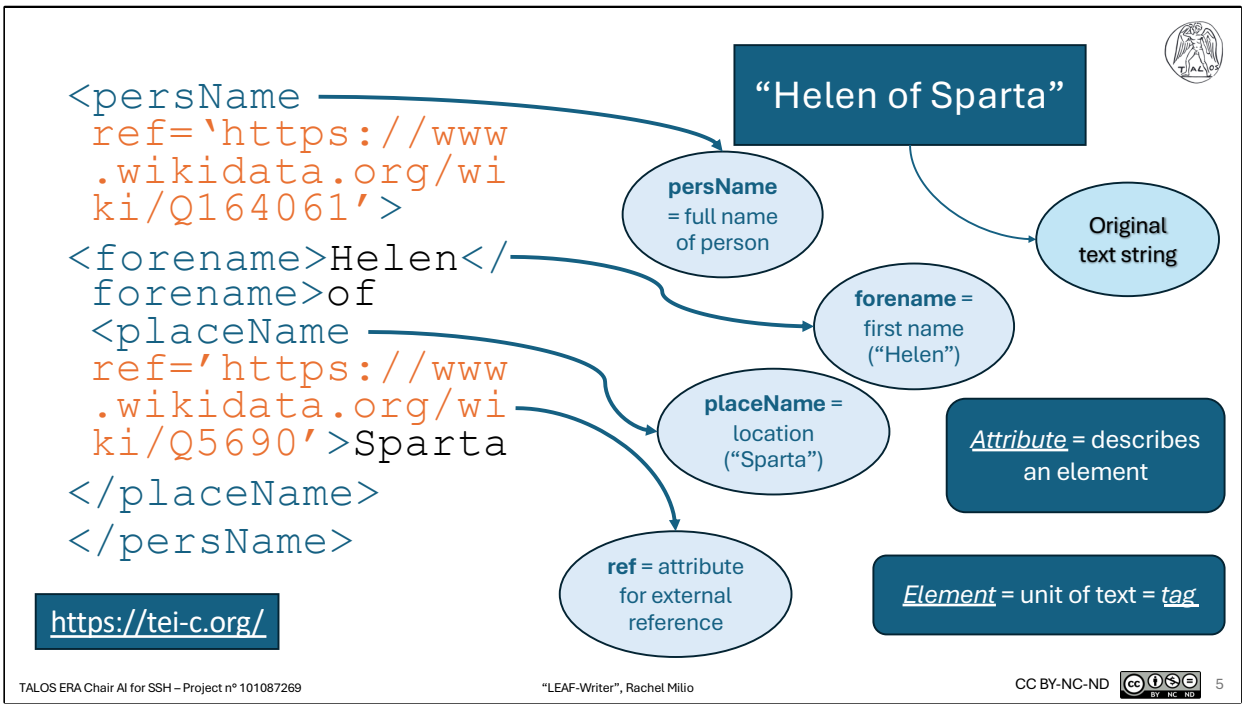
First, it is necessary to introduce the field of text encoding. Through text encoding, we can make human-readable texts machine-readable. Another term for text encoding is *mark-up*. There are three main types of text mark-up: structural, presentational, and semantic. Structural markup denotes the structure 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).



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.



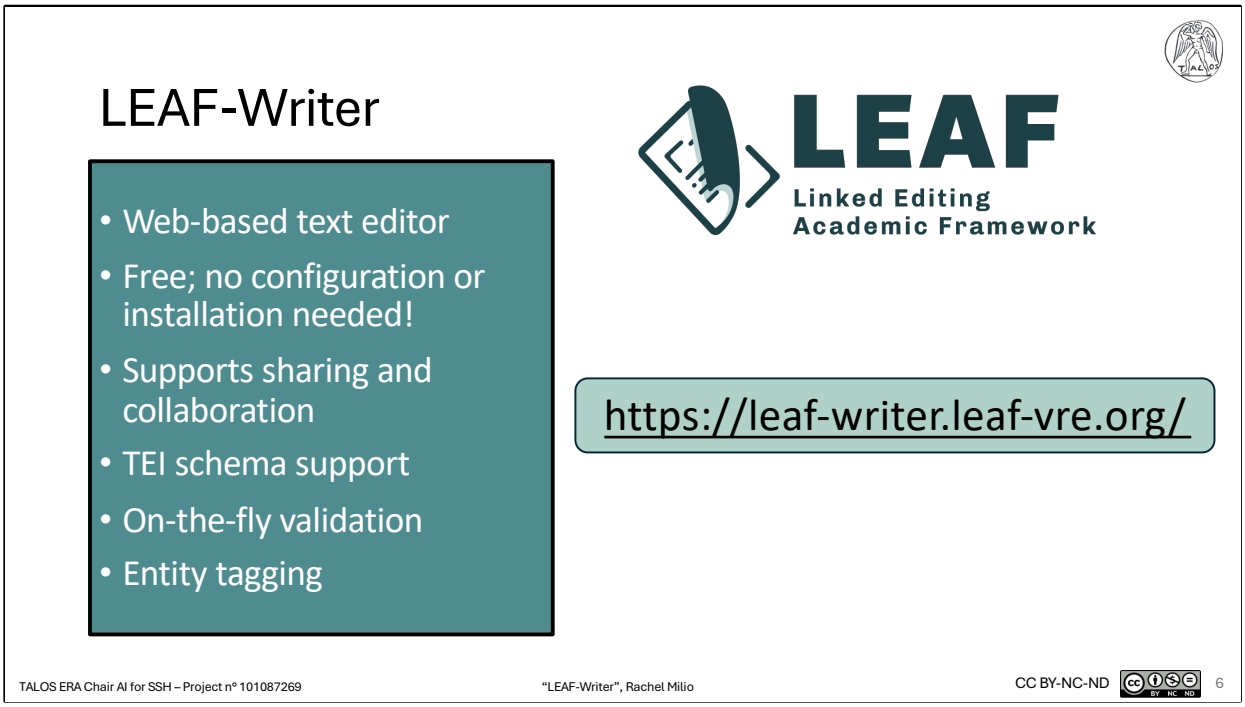
TEI, or the Text Encoding Initiative, is a humanistic standard for XML text encoding. Through the use of TEI, text encoding can be more consistent, increasing interoperability and reusability. TEI is most commonly used for the digital encoding of literary and linguistic texts, such as manuscripts, historical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



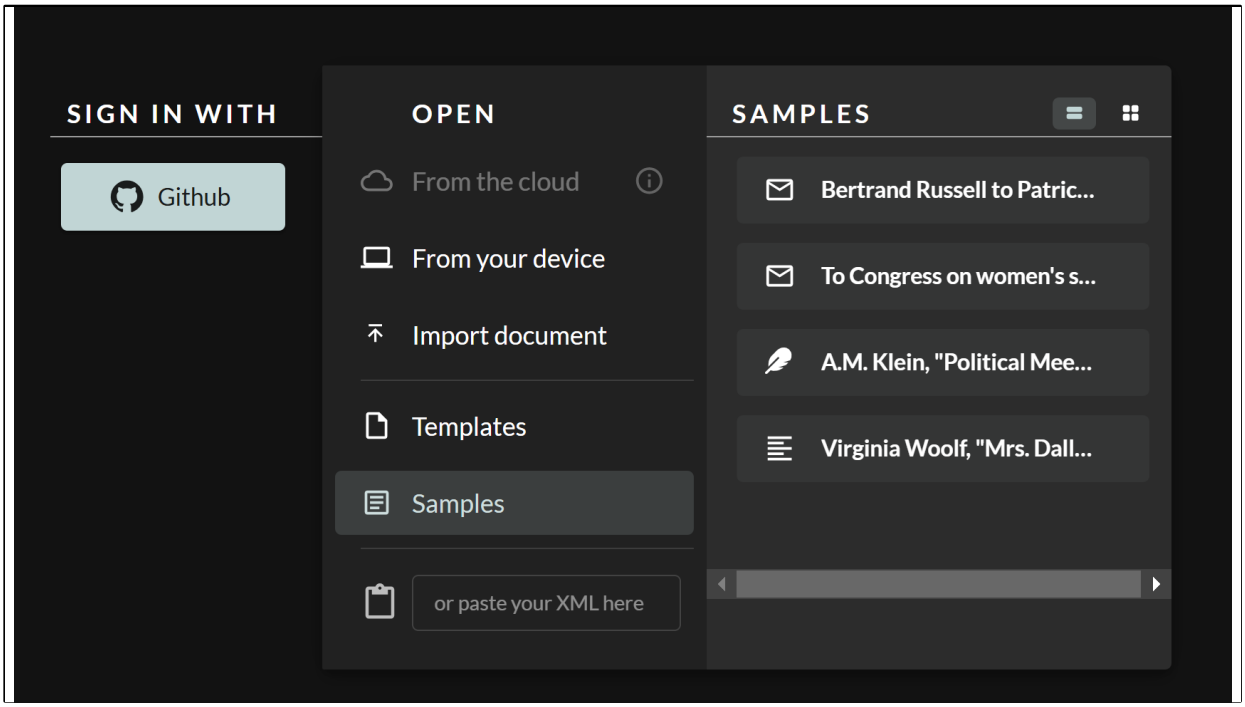
In TEI, an element is a unit of text. These are also known as tags. An attribute describes an element. For example, the sample text "Helen of Sparta" can be tagged as follows:

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

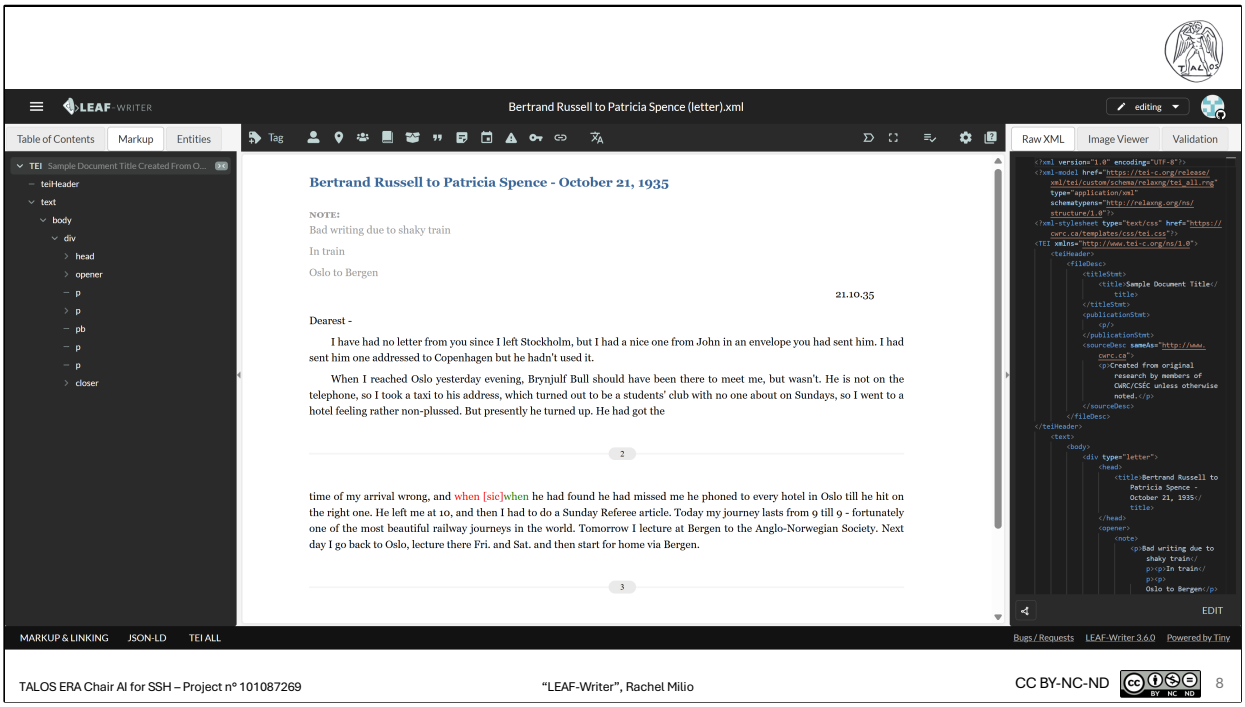
In this case, the nested tags designate the entire string "Helen of Sparta" as the name of a person, while "Helen" is a forename and "Sparta" is a place name. The attribute "ref" is used to point to an external URI, in this case from the authority base Wikidata. For more on TEI, I recommend checking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



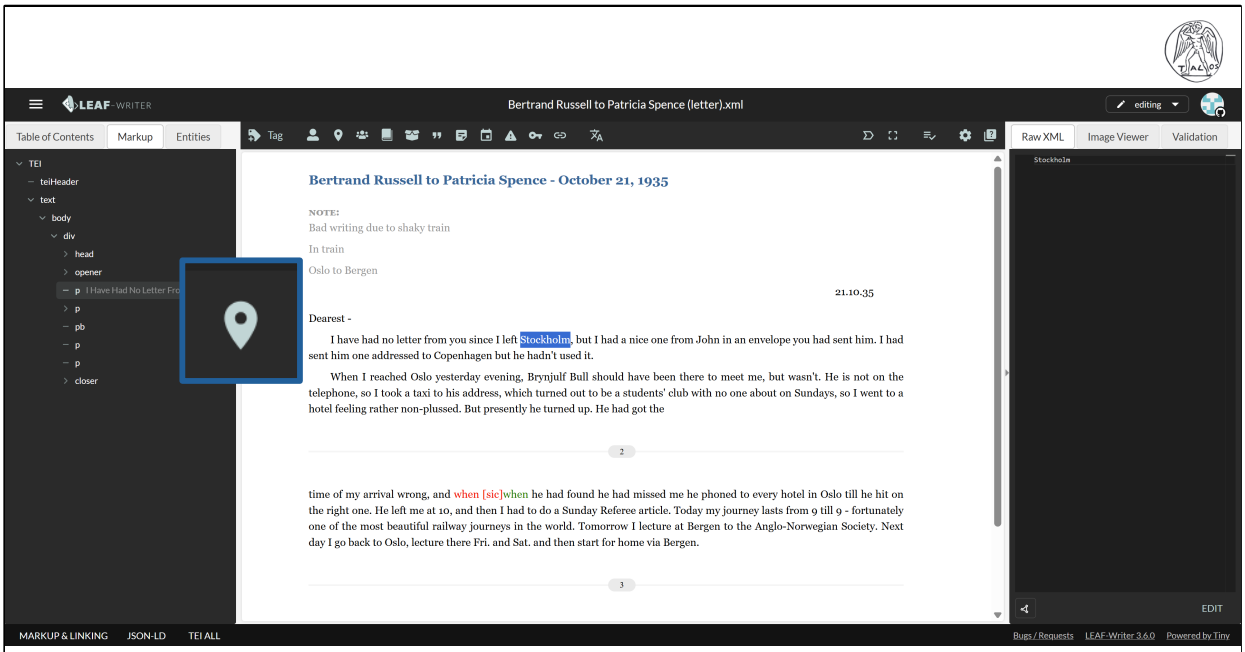
Now that we understand what TEI is and how it is used in the digital humanities, we can introduce the platform LEAF-Writer. LEAF-Writer is a free, web-based semantic text editor. Unlike other XML editors, many of which are proprietary, LEAF-Writer is completely open-source and requires no configuration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



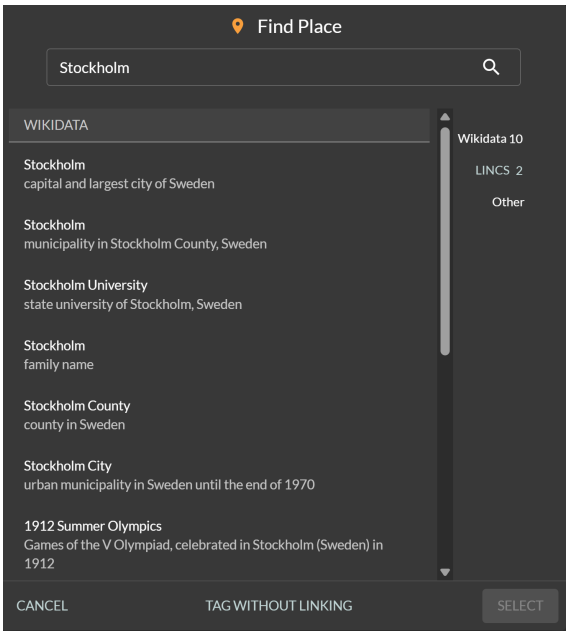
On the LEAF-Writer landing page, you can see the options to open a document. By logging in to Github, you can access documents stored via cloud, but you can also open documents stored locally on your computer or import from the OCR site Transkribus. LEAF-Writer also provides templates and sample documents 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.

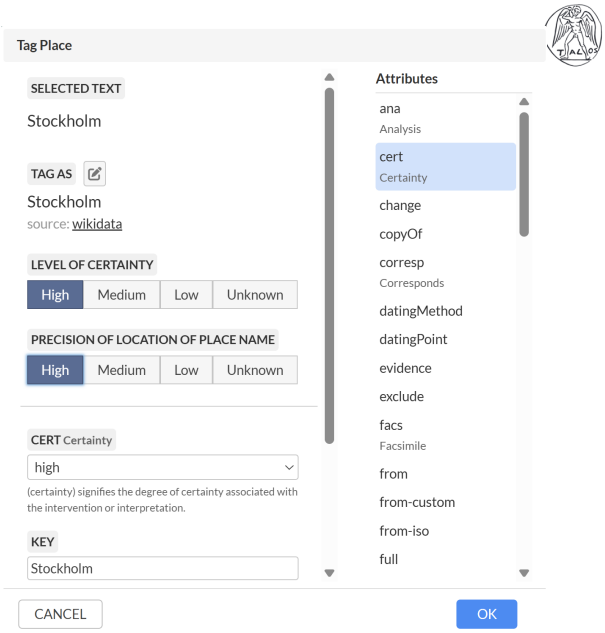


This is the editing view of LEAF-Writer. On the left-hand side, we have three tabs: the Table of Contents, the Markup, and the Entities. The "table of contents" displays section headers (for example, chapter titles in a longer text). "Markup" displays the TEI-XML elements in their nested structure. "Entities" 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.

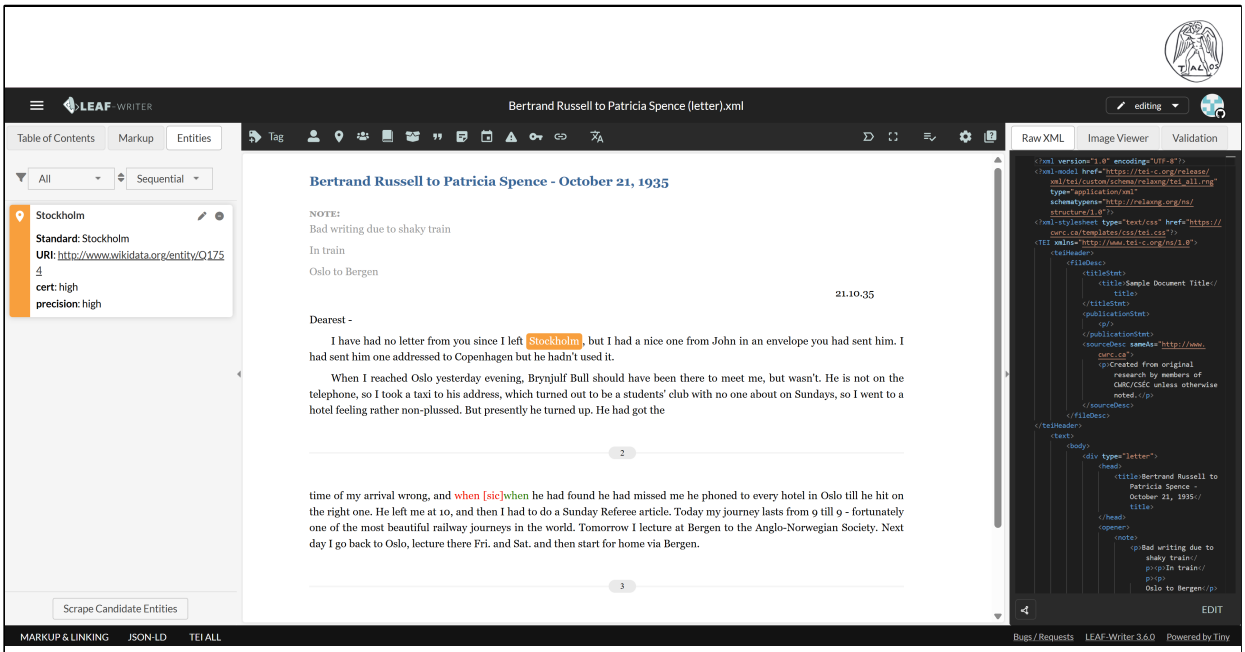


Let's try out tagging an entity. If we highlight the text "Stockholm" and select the Place icon, you'll be prompted by a pop-up to select the correct place identifier.

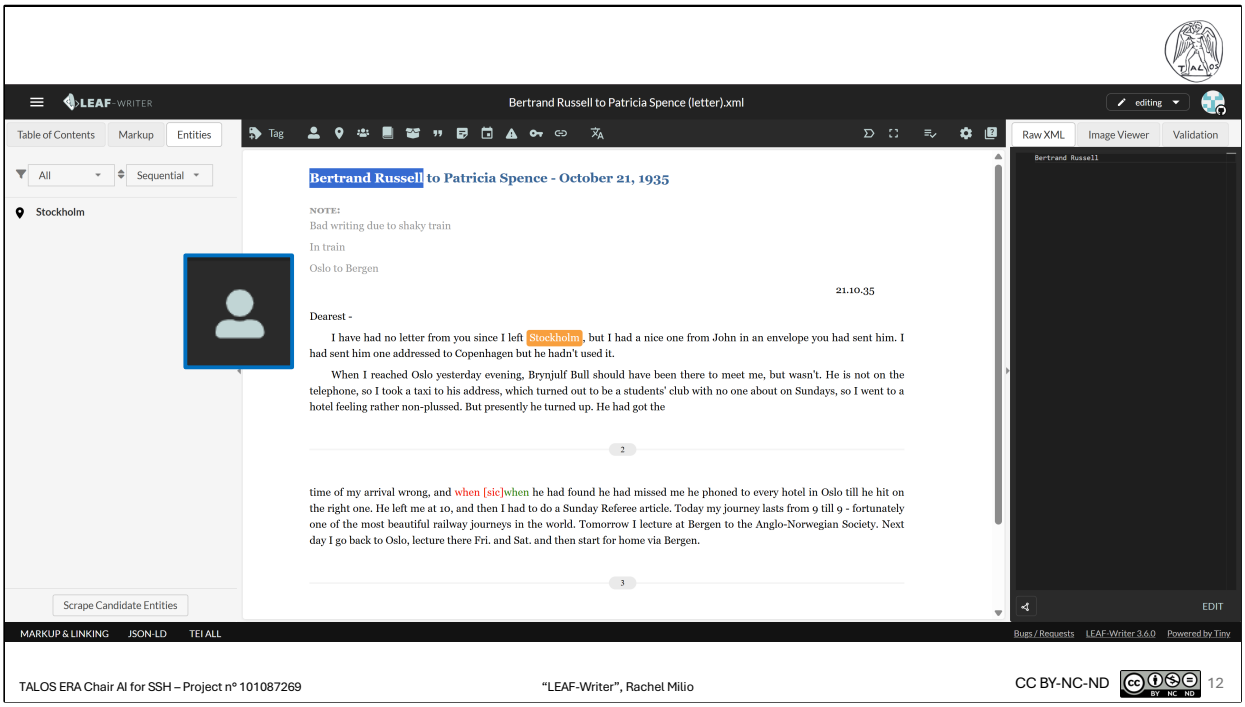




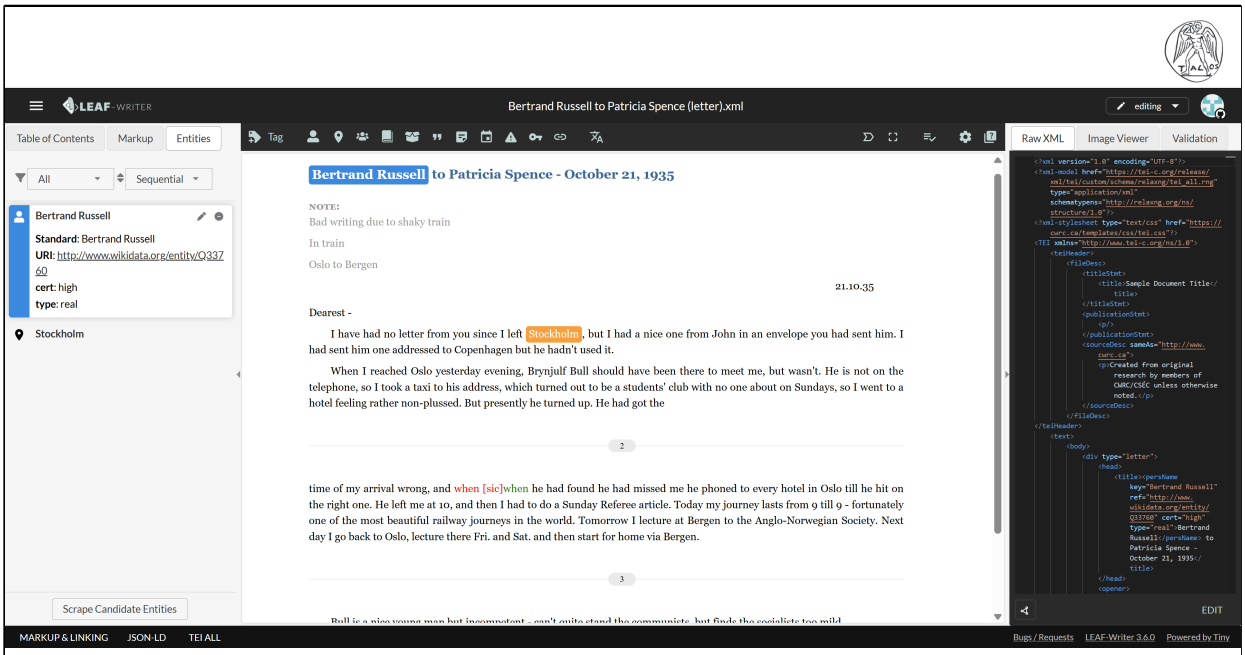
The text "Stockholm" refers to the Swedish capital city Stockholm, so we'll select that option. Now you can select attributes such as certainty and precision.



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

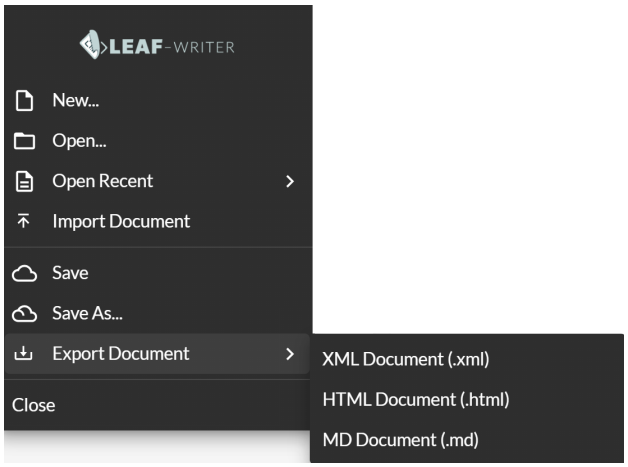


Next, highlight the name Bertrand Russell in the title. Select Person, and choose the appropriate identity.



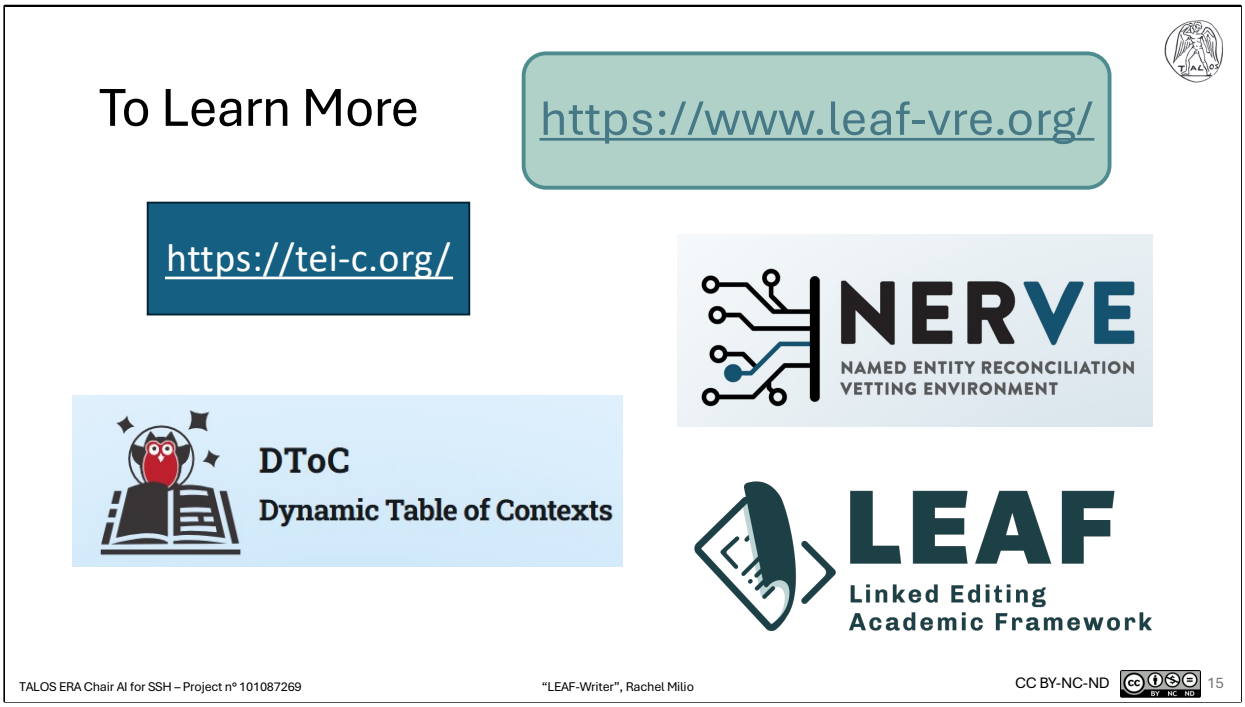
Now, you've tagged the string "Bertrand Russell" as a person entity.

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.



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

https://talos-ai4ssh.uoc.gr/

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

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