EXERCISE 03 Named entities - INSTRUCTIONS:

TOOL (Creating a new TEI-XML File)

- Open oXygen (to open, click on "oXygen XML Editor").
- Open new document (oXygen menu: File -> New document).
- Search or select a document type (Framework templates -> TEI P5 -> TEI All [TEI P5]).
- Save it and rename it.
- Tip! You can save it in the same folder of this exercise.

TRANSCRIPTION

- Not need to transcribe (no time for that now), just have a look at the transcription provided "3.1.Stendhal_Memoires_1838_transcript.txt". The text has 3 paragraphs, so we need 3 `<p>` elements.
- Remove the dummy text "Some text here." between the opening `<p>` and the closing `</p>` within the `<body>` element. Add 2 more `<p>` elements (opening and closing).
- Copy and paste each paragraph from the transcription.
- There are other ways of doing all this: copy and paste the entire text within a `<p> </p>`, put the cursor at the beginning of each paragraph, go to Document -> Markup -> Split Element.

The document should be now well-formed (a green square is visible in the upper right-hand corner).

- Tip! To format and indent the encoding, press Cmd + Shift + P (Mac), Ctrl + Shift + P (Windows).

ENCODING

You are encoding in the `<body>` names of places, names of people, and dates, so you need to add these elements:

- `<placeName>`, `<persName>` and `<date>`.
- Tip! Remember the shortcut: Highlight the place (person or date), press Ctrl + E (Windows) or Cmd + E (Mac), start typing 'pl..' and press ENTER after selecting your element.

Let's add now the metadata about the named entities (places and persons) in the `<teiHeader>`. After this is done, you will be able to link this information to each element in the text. For now, you need to add these elements:

- `<profileDesc>` describes non-bibliographic aspects of a text.
  - `<particDesc>` persons named in the text.
    - `<listPerson>` list of persons.
      - `<person>` information about a person.
  - `<settingDesc>` places named in the text.
    - `<listPlace>` list of places.
      - `<place>` information about a place.
- Start adding `<profileDesc>` just after `</fileDesc>` and before `</teiHeader>`.
- Tip! Place the cursor right after `</fileDesc>` (or at the next line) and start typing `<` (angle bracket): oXygen suggests a list of element names available at this point.
- Add the elements `<particDesc>` and `<settingDesc>` within `<profileDesc>`.
- Add the element `<listPerson>` within `<particDesc>`.
- Add the element `<listPlace>` within `<settingDesc>`.

Don't worry about the red square, some mandatory elements are still missing. Add `<person>` within the element `<listPerson>`, and `<place>` within the element `<listPlace>`. The structure should now be well-formed and look like:

```
<profileDesc>
   <particDesc>
      <listPerson>
        <person></person>
      </listPerson>
   </particDesc>
   <settingDesc>
      <listPlace>
        <place></place>
      </listPlace>
   </settingDesc>
</profileDesc>
```

There are 4 places and 4 person names in the text. For now let's gather the information about one from each kind: Paris and Tocqueville.

Encoding places

Let's start with **Paris**. We are interested in encoding its location in terms of latitude and longitude. To obtain coordinates you can search a geographical database (gazetteer), e.g, GeoNames: https://www.geonames.org/2988507. You should end up collecting the following information, which you are going to encode in the element `<place>`.

place name	*coordinates*	*geonames-id*	*country*	*your own id*
Paris	48.85341,2.3488	2988507	France	paris

Within the element `<place>` you need to add the following elements and complete the information from the table:

- `<placeName>`
- `<country>`
- `<location>`
  - `<geo>` (coordinates separated by a comma, inside `<location>`)
- `<idno>` (a standardized id, in this case from GeoNames)

Unique identifier. In order to link `<place>` from the header to the element `<placeName>` in the text, we need to specify a **unique identifier** for the element you just have created. We use attributes for this:

- Add the attribute `@xml:id` to the element `<place>` and choose your own unique value, e.g., "paris". You should have now `<place xml:id="paris">`.
- Tip! Put the cursor inside the element just before `>` (`<place↑>`), and press space; oXygen suggests a list of attributes available for this element.
- `<idno>` This element needs also an attribute and a value to supply the information, e.g.: `type="GeoNames"`.

Internal linking. Now that you have defined the place, in this case, Paris, and supplied a unique identifier, you can link this element with the all place names in your text, which refer to Paris:

- To do the linking, add the attribute `@ref` containing the value of the `@xml:id` (i.e., "paris") preceded by a hash (`#`). You should have now `<p>À <placeName ref="#paris">Paris</placeName>,`

`la distraction est trop [...]`.

- Tip! Put the cursor inside the element just before `>` (`<placeName↑>`), and press space; oXygen suggests a list of attributes available for this element. Choose `@ref`: automatically oXygen will suggest the identifiers you have created in the header.

Encoding persons

Let's continue with a person name, **Tocqueville**. Wikipedia may be there always to help, but a good source for the basic infos are the authority files run by national libraries, e.g., the data service of the *Bibliothèque nationale de France*, https://data.bnf.fr. The international authority files, like INSI http://www.isni.org/isni/0000000121371933 gather several of these sources.

So once you collected the infos you want to add, encode them in the element `<person>`.

person name	*birth*	*death*	*occupation*	*nationality*	*your own id*
Alexis de Tocqueville	29 July 1805	16 April 1859	historian	French	AT1805

ISNI id
0000000121371933

Within the element `<person>` you need to add the following elements and complete the information from the table:

- `<persName>`
- `<birth>`
- `<death>`
- `<occupation>`
- `<nationality>`
- `<idno>` (a standardized id, in this case ISNI).

Unique identifier. As we did with `<place>` we need to specifice a **unique identifier** for `<person>`. It is up to you, which want to use, e.g., the initials plus the date of birth: AT1805 (**A**lexis de **T**ocqueville **1805**).

- You should have now `<person xml:id="AT1805">`.

You probably added the date of birth like `<birth>29 July 1805</birth>`. In fact you can write it as you prefer: 29 juillet 1805, 29.07.1805, etc., but in order to record the date properly, we need to use the attribute `@when` with a value in a standard form: 1783-01-23. So you should have `<birth when="1783-01-23">29 July 1805</birth>`.

- Keep tagging all dates with the attribute `@when`, also in your text.

Internal linking. Now that you have defined the person, in this case, Tocqueville, and supplied a unique identifier, you can link this element with the names in your text, which refer to Tocqueville:

- Proceed in the same way you did with with `<place>` and remember the tip! of putting the cursor inside the element (or inside the value attribute).

Already done?

- (1 option) Search for infos about Charles Louis de Montesquieu, Pierre Bayle, Lyon, Marseille, Grenoble, and keep tagging... Not your thing?
- (2 option) Could you improve the encoding of `<persName>` with forename and surname?

- (3 option) You could also add to the person list the information about the author, Stendhal, which is in fact a pseudonym of his real name Marie-Henri Beyle. How do we encode a pseudonym? Check the TEI Guidelines for 13.2.1 Personal Names.
- (4 option) Visualize your data. Anything fancy here, but you can spot with ease the names, places and dates.

VISUALIZATION (Web browser and oXygen editor mode)

- Follow the steps as in exercise 1.

EXERCISE SOLUTION

- I provided you with a TEI-XML file already encoded, "3.4-Stendhal_Memoires_1838_done.xml". Feel free to have a look inside or ask me for help.