XML-TEI basis

Ariane Pinche

Warsaw, May 20–24, 2019.

1. What is a digital edition?

1.1 What is a text for computers?

- characters string
- fonts : https://folk.uib.no/hnooh/mufi/fonts/

1.2. What is a text for humans?

- A string of characters that most of the time makes sense.
- A text has different levels of hierarchy:
    - grammatical
    - structurel
    - semantic

1.3 Adding information in a text

There are two kinds of encoding:

- layout encoding:
    - *italic*, **bold**, etc.
- semantic encoding
    - language, type of word, etc.

Langage	Layout encoding	Semantic encoding
LaTeX	emph{ad hoc}	\selectlanguage{latin}{ad hoc}
HTML5	`<i> ad hoc </i>`	`<i lang="la">ad hoc</i>`
XML-TEI	`<hi rend="i">ad hoc</hi>`	`<foreign xml:lang="la">ad hoc</foreign>`

1.4 XML solution

http://shelleygodwinarchive.org/sc/oxford/frankenstein/volume/i/#/p1/mode/rdg

XML can be used:

- to propose a representation of a text
- to make the scientific edition
- to allow access the text in its linearity and also as a database to interrogate.

2-What is XML

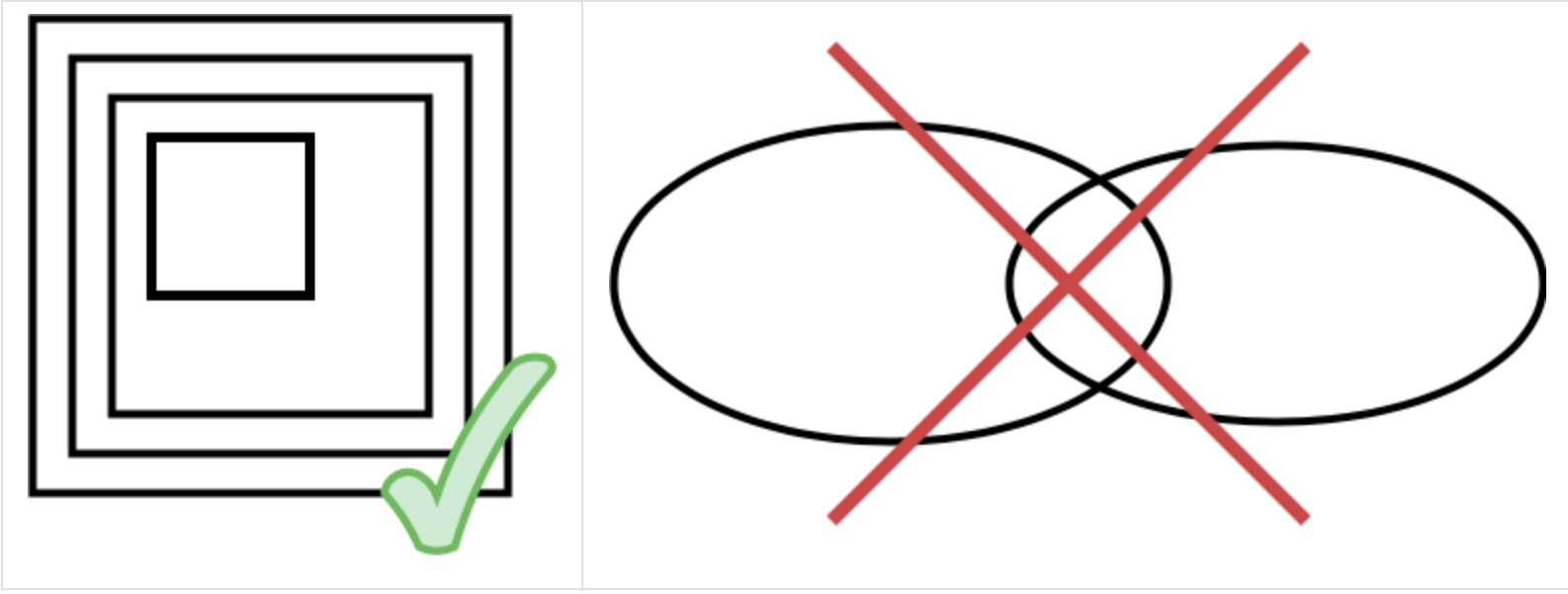
2.1 XML definition

eXtensible Markup Language

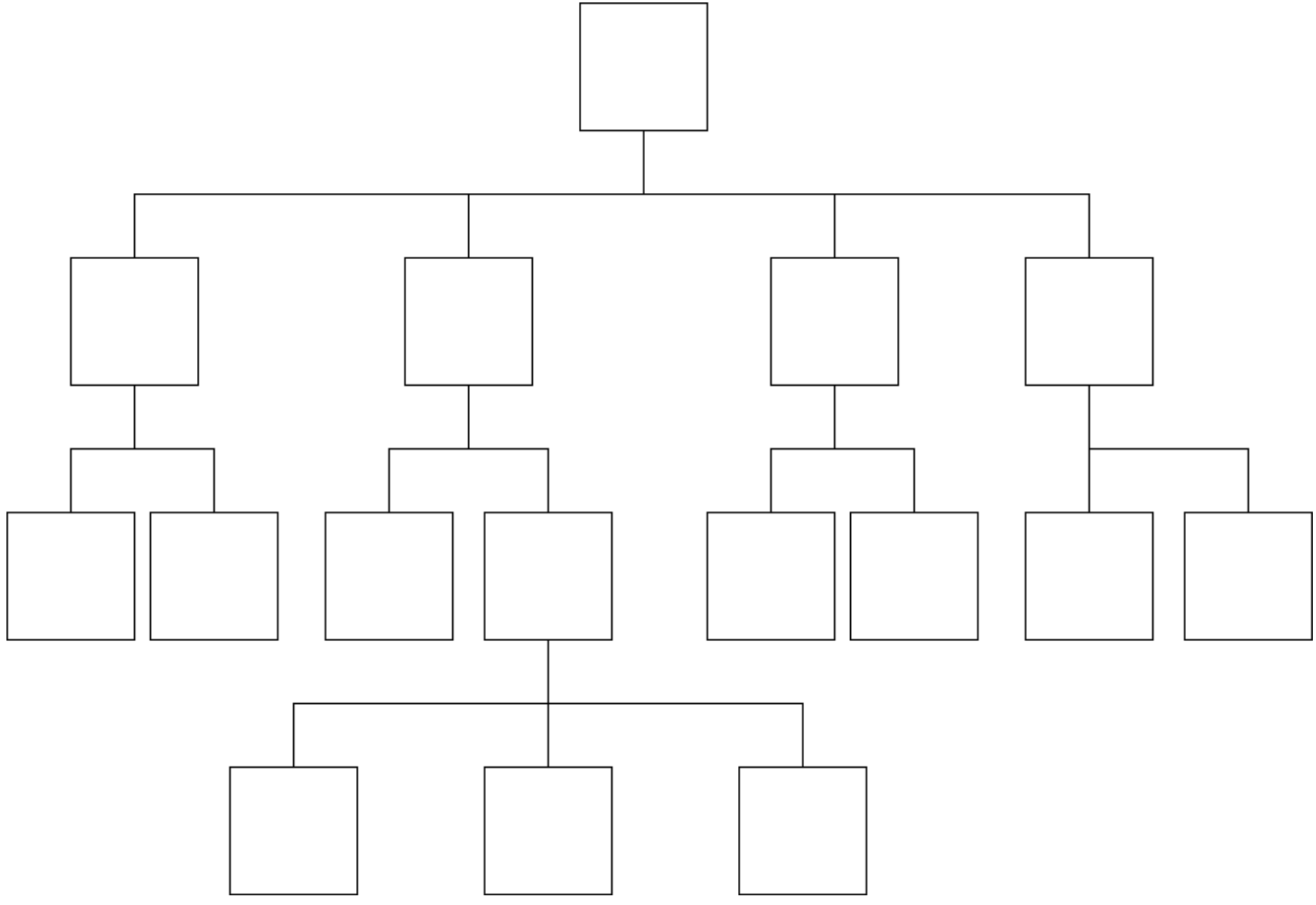
- XML is an international language since 1998 following the W3C recommendations.
- It is data format really easy and well documented.
- XML is a free language in the middle of a large community.
- XML was designed to store and transport data easily
- XML was designed to be both human- and machine-readable.
- It is a markup language created for text.
- XML has non-predefined markup set, this is why it is extensible.

2.2 XML structuration

- XML has hierarchical construction that works on imbrication.



- XML is a tree with a root element, every other elements are its descendant



- XML markups are elements:

```
<ElementName>
textString
</ElementName>
```

- An element can have an attribute or multiple attributes

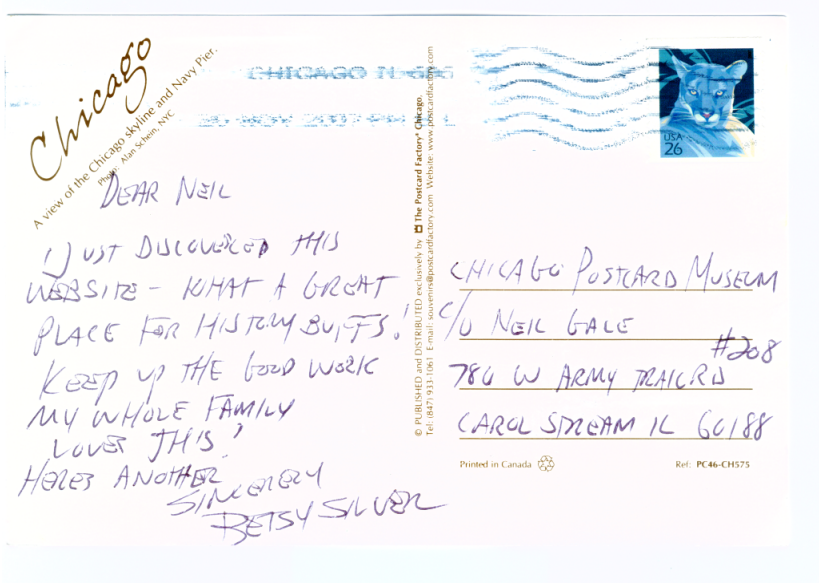
```
<ElementName nameAttribute='attributeValue'>
textString
</ElementName>
```

- XML elements work on the principle of imbrication:

```
<Element1>
  <Element2>
      textString
  </Element2>
</Element1>
```

2.3 Encoding

Try to encode a post card



- Create an XML element to encode the complete object
- Create an XML element to encode the recto side
- Create an XML element to encode the verso side
- On the verso side :
    - Create XML elements to encode the heading
    - Create XML elements to encode the text
    - Create XML elements to encode the address

3-What is TEI?

3.1 Definition

Text Encoding Initiative

- TEI is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.
- TEI is a dataset of XML mark-up to help describe a text in digital form.

3.2 What is the interest of TEI?

- Meaning of the text before format;
- Software independence;
- Community-driven;
- Standard to exchange
- Data curation.

3.3 A bit of History

• 1987: birth of Text Encoding Initiative.
• 1990: TEI P1 (proposal 1), dir. Michael Sperberg-McQueen et Lou Burnard.
• 1992–1993: TEI P2, expansion.
• 1994: [TEI P3] (http://www.tei-c.org/Vault/GL/P3/index.htm), first complete version.
• 2000: birth of TEI Consortium.
• 2001–2004: TEI P4 uses XML.
• 2007— ...: [TEI P5] (http://www.tei-c.org/Guidelines/P5/), stop usage of SGML.

3.4 How to use it?

Translate the postcard XML markup in TEI wiht the following tags :

- text, body, div, p
- span
- closer, salute, signed
- address, persName, street, settlement

3.4.1 Base structure of a TEI document

1. An instruction XML
2. Schema declaration
3. Root element
4. Two mandatory subelements:
    - teiHeader
    - text

- teiHeader is for metadata of the document: title / author / type of encoding, etc.

```
<teiHeader>
    <fileDesc>
       <titleStmt>
          <title>Title</title>
       </titleStmt>
       <publicationStmt>
          <p>Publication Information</p>
       </publicationStmt>
       <sourceDesc>
          <p>Information about the source</p>
       </sourceDesc>
    </fileDesc>
```

- teiHeader can have 4 subsections:
    - fileDesc (obligatory)
    - encodingDesc
    - profileDesc
    - revisionDesc
- fileDesc has 3 obligatory parts:
    - titleStmt
    - publicationStmt
    - sourceDesc

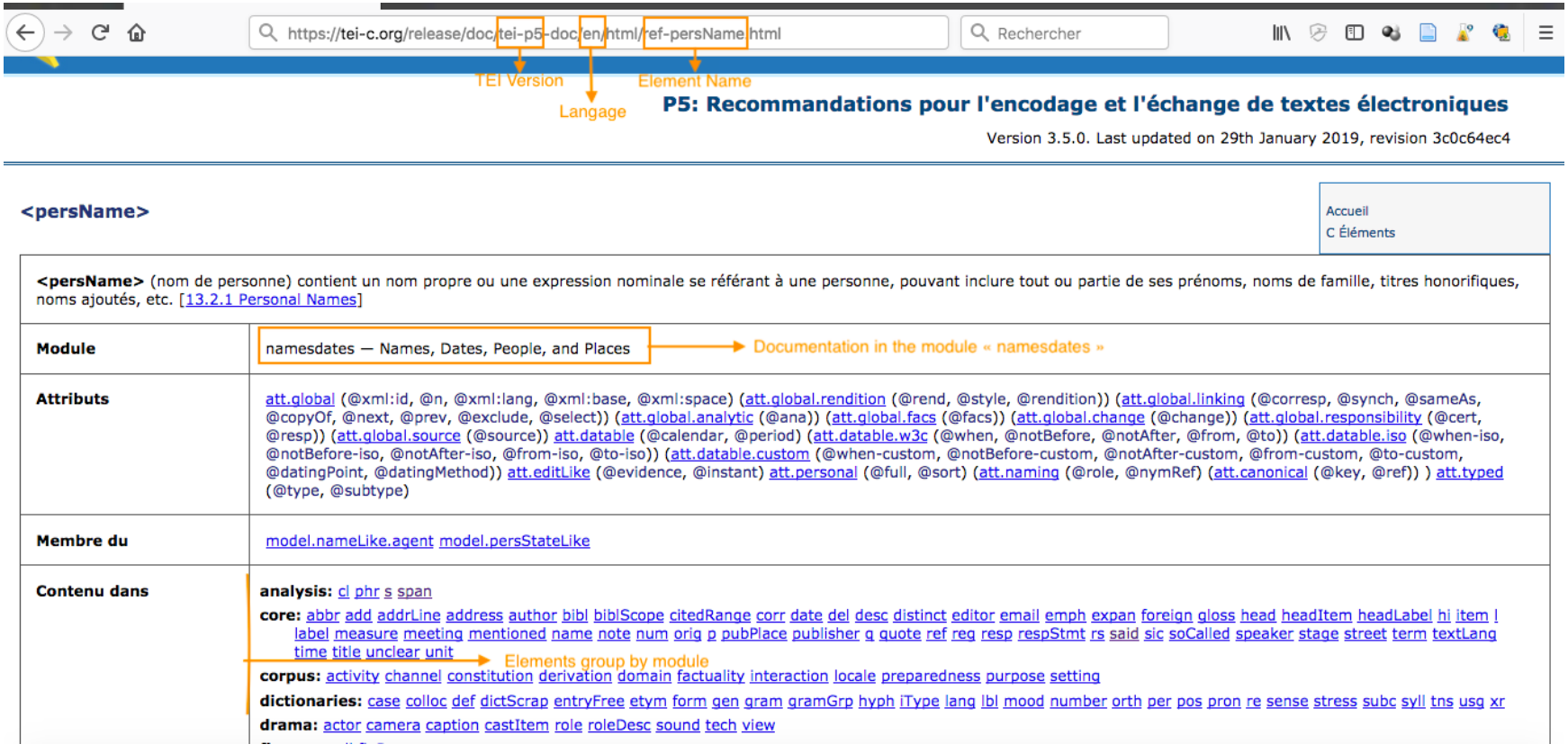
- text is the markup where you put your textual material, you can arrange it in 3 parts:
    - front
    - body
    - back

3.4.2 Documentation: how to read the guidelines

3.4.2.1 Develop documentation

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

3.3.2.2 Documentation for a specific element



Exercice

With the guidelines help try to structure the “Sacramenta Argentariae” (The Oaths of Strasbourg)

add :

- paragraphs
- speech and information about languages used in the speech
- persName tags
- simple teiHeader

Transcription : https://fr.wikisource.org/wiki/Serments_de_Strasbourg