

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	<code>emph{ad hoc}</code>	<code>\selectlanguage{latin}{ad hoc}</code>
HTML5	<code><i> ad hoc </i></code>	<code><i lang="la">ad hoc</i></code>
XML-TEI	<code><hi rend="i">ad hoc</hi></code>	<code><foreign xml:lang="la">ad hoc</foreign></code>

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

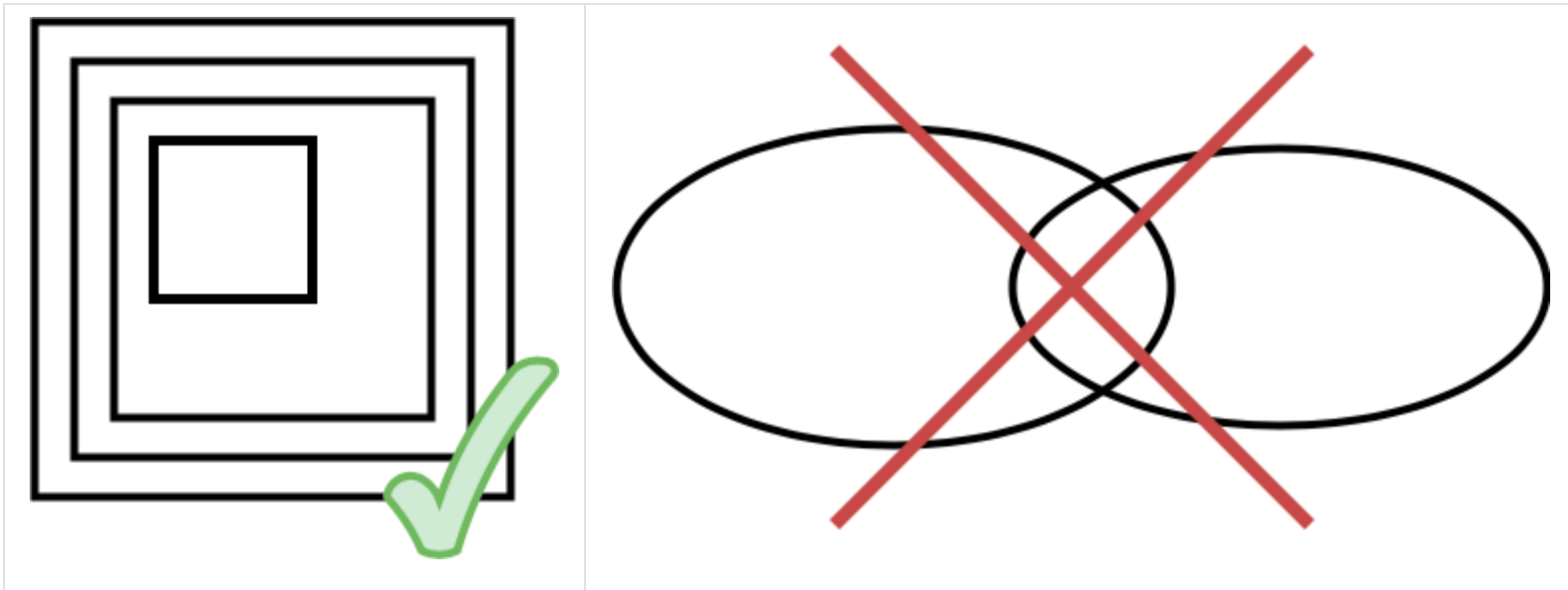
2.1 XML definition

eXtensible Markup Language

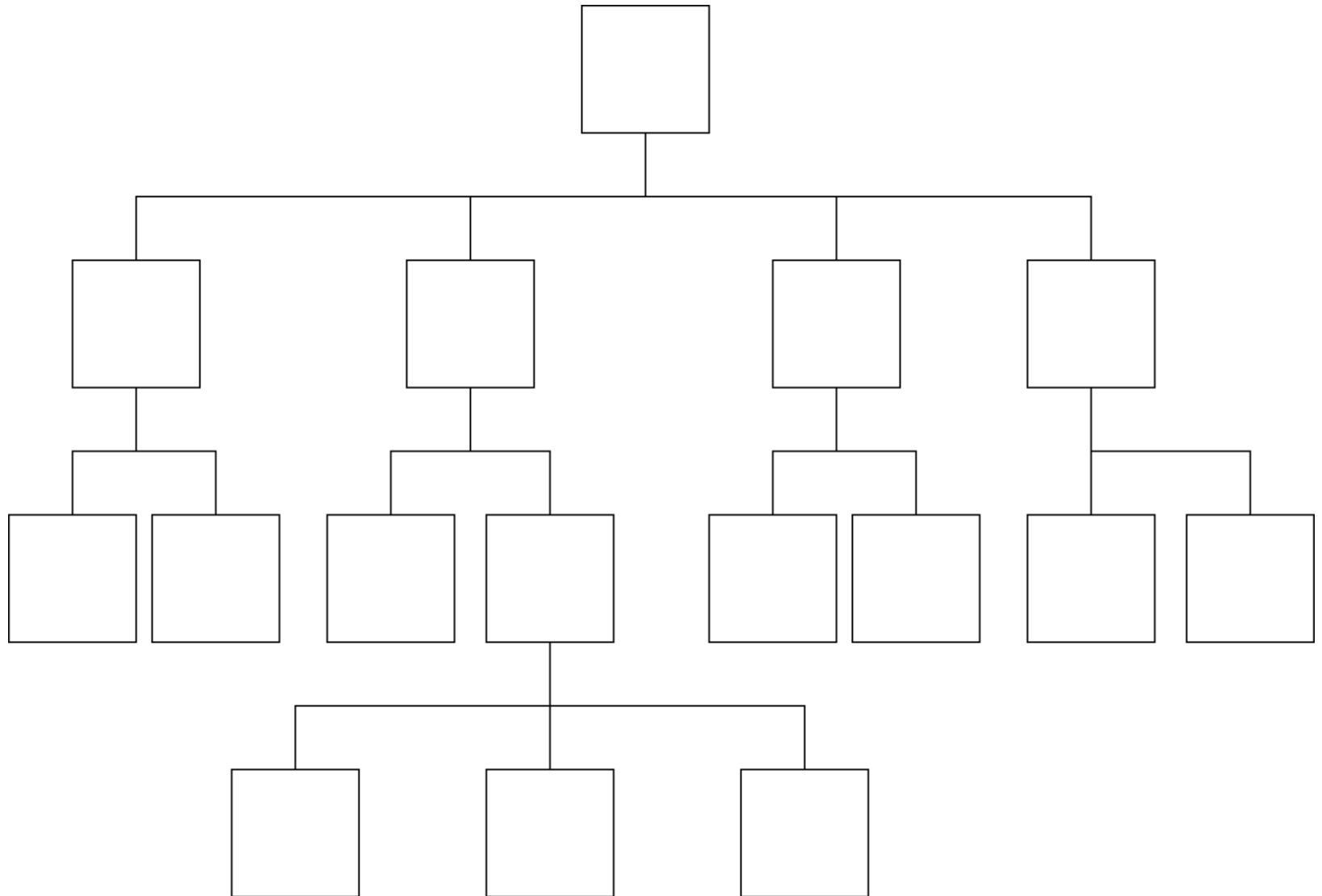
- XML is an international language since 1998 following the W3C recommendations.
- It is data format really easy and well documented.
- XML is a free language in the middle of a large community.
- XML was designed to store and transport data easily
- XML was designed to be both human- and machine-readable.
- 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>
```

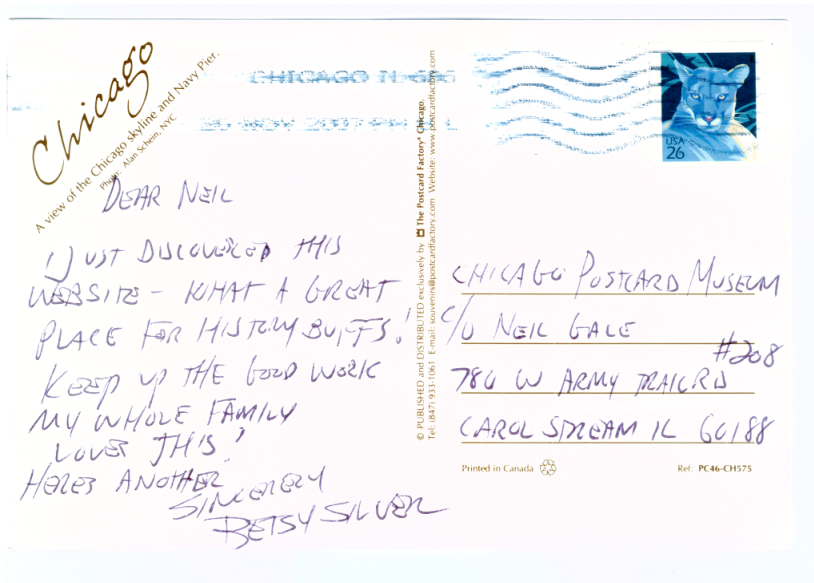
- XML elements work on the principle of imbrication:

```
<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

[←](#)
[→](#)
[↺](#)
[🏠](#)

TEI Version

Language

Element Name

P5: Recommandations pour l'encodage et l'échange de textes électroniques

Version 3.5.0. Last updated on 29th January 2019, revision 3c0c64ec4

<persName>

Accueil

C Éléments

<persName> (nom de personne) contient un nom propre ou une expression nominale se référant à une personne, pouvant inclure tout ou partie de ses prénoms, noms de famille, titres honorifiques, noms ajoutés, etc. [\[13.2.1 Personal Names\]](#)

Module	<div>namesdates — Names, Dates, People, and Places</div> <div>Documentation in the module « namesdates »</div>
Attributs	att.global (@xml:id, @n, @xml:lang, @xml:base, @xml:space) (att.global.rendition (@rend, @style, @rendition)) (att.global.linking (@corresp, @synch, @sameAs, @copyOf, @next, @prev, @exclude, @select)) (att.global.analytic (@ana)) (att.global.facs (@facs)) (att.global.change (@change)) (att.global.responsibility (@cert, @resp)) (att.global.source (@source)) att.dateable (@calendar, @period) (att.dateable.w3c (@when, @notBefore, @notAfter, @from, @to)) (att.dateable.iso (@when-iso, @notBefore-iso, @notAfter-iso, @from-iso, @to-iso)) (att.dateable.custom (@when-custom, @notBefore-custom, @notAfter-custom, @from-custom, @to-custom, @datingPoint, @datingMethod)) att.editLike (@evidence, @instant) att.personal (@full, @sort) (att.naming (@role, @nymRef) (att.canonical (@key, @ref))) att.typed (@type, @subtype)
Membre du	model.nameLike.agent model.persStateLike
Contenu dans	<p>analysis: cl phr s span</p> <p>core: abbr add addrLine address author bibl biblScope citedRange corr date del desc distinct editor email emph expan foreign gloss head headItem headLabel hi item label measure meeting mentioned name note num orig p pubPlace publisher q quote ref req resp respStmt rs said sic soCalled speaker stage street term textLang time title unclear unit → Elements group by module</p> <p>corpus: activity channel constitution derivation domain factuality interaction locale preparedness purpose setting</p> <p>dictionaries: case colloc def dictScrap entryFree etym form gen gram gramGrp hyph iType lang lbl mood number orth per pos pron re sense stress subc syll tns usg xr</p> <p>drama: actor camera caption castItem role roleDesc sound tech view</p>

Exercise

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