UNIVERSITÀ di VERONA
Dipartimento di LINGUE E LETTERATURE STRANIERE
Scuola di dottorato in SCIENZE UMANISTICHE

Dipartimento di Eccellenza 2018/2022 MIUR

Digital Scholarly Editions and XML TEI markup language

PROGETTO MAMBRINO

Stefano Bazzaco
stefano.bazzaco.1@gmail.com

ABCDH

READ coop

Transkribus

EnExDi2022
University of Poitiers – 10/05/2022

Seminar structure

- Introduction to Digital Scholarly Editing and Digital Scholarly Editions
- Textuality in the Digital Realm
- At the core of DSE: data and model
- XML TEI as a descriptive markup language
- Introduction to XML TEI documents

DIGITAL SCHOLARLY EDITING

One of the most developed fields inside DH community

It refers to the act of creating a **Digital Scholarly Edition (DSE)**

Digital: produced in a digital environment

Edition: the representation of a text (mainly referring to literary text)

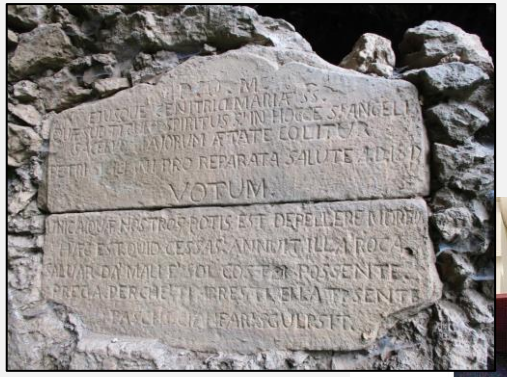
Scholarly: (read: «scientific») created starting from a set of rules, as to say a documented methodology

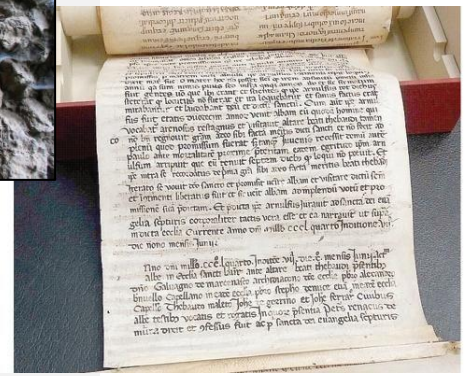
Does it changes anything when the editorial process migrates to the digital realm?

Has the medium some transformative power on the objects it deals with?

TEXTUALITY (1)

starting point: texts in the Analog Realm





the concepts of **document – writing – textuality** have been seen for centuries as something INSEPARABLE

→ to texts it has been reserved **the production, transmission and preservation of knowledge** (independently from the medium: from inscriptions to rotulos; from codexes to printed books)

TEXTUALITY (2)

The migration of texts to the computational field put the evidence on the fact that texts are **composite objects** (symbolic nature of texts)

Each aspect of the text constitutes an **informative layer** (f.e. form/content, materiality/meaning)

Combined together, all these layers constitues a unique artefact that is the result of a **mediation** (not trivial) between expression and medium

TEXTUALITY (3)

In the Analog Realm, texts are perceived as a single phenomenon resembling a set of different characteristics

> it is an inner AMBIGUITY that in human perception does not obstaculize its usage and function for readers

Text as:

- **the artefact (document, book, …)**
- **the linguistic content**
- **but also: verbal content, set of graphemes/phonemes, group of pages, …**

each of them is an informative layer

TEXTS IN THE DIGITAL REALM

Computers require a high level of **FORMALIZATION**:

a) each informative layer of a text has to be taken into account independently;

b) the conjunction of them produces an **COMPLEX informative system** comparable to the book in the Analog Realm

c) the migration of texts to a digital environment deals with:

- computers abstraction capabilities
- the historic and interpretative tradition of modelled contents
- the creation of a complex object

FROM TEXT TO DATA

The formalization and migration of every layer provides a set of data

DATA stays at the core of DSE (and is obtained by an act of ***digitalization***):

1. the transcription of the content in MRF (*machine readable form*)
2. the photographic reproduction of a text (from materiality to images)
3. other data: bibliographic databases, audio files, maps,…

FROM DATA TO DSE

The organization of data into a system

- following a **conceptual model**
- respecting some **methodology**
- and including **hierarchies and functionalities**

leads to the creation of a Digital Scholarly Edition

Depending on:

- **which kind of data** is included (transcriptions, images, other materials,…)
- **how the data is created, combined and displayed**

we can distinguish between different types of digital editions

TEXT ENCODING

To encode a text means to «translate» it for the computer

Digital editing phases:

- Analysis of text to enstablish an encoding model
- Choosing of an encoding language
- Creation of an encoding schema
- **First encoding act: convert the text in MRF**
- **Second encoding act: markup**
- Application of stylesheets / ODD schemas (visualization)
- Publication and distribution

TEXT ENCODING (2)

Texts can be considered as encoded:

- semiotic/linguistic level
- presentational level (characters encoding)
- material level (layout)

I PROMESSI SPOSI.

CAPITOLO PRIMO.

uel ramo del lago di Como, che volge a mezzogiorno, tra due catene non interrotte di monti, tutto a seni e a golfi, a seconda dello sporgere e del rientrare di quelli, vien, quasi a un tratto, a ristringersi, e a prender corso e figura di fiume, tra un promontorio a destra, e un'ampia costiera dall'altra parte; e il ponte, che ivi congiunge le due rive, par che renda ancor più sensibile all'occhio questa trasformazione,

Modelling means to point out all these features in a way that the machine can understand them

MODELLING

To model the digital text is an essential phase:

The model determines what to markup inside the text:

- *Linguistic features* (phonemes, morphemes, sintactic units,…)
- *Stilistic features* (stylemes, rethoric figures,…)
- *Presentational features* (what to retain of the material source)
- *Research features* (mark up the research specific contents, f.e. names, dates)

A model must include **only relevant features for the project**

MODELLING (2)

In other words:

- **A text is much more than a sequence of characters**
- By markup we make its inner features explicit
- Only explicit markup can be interpreted and elaborated by computers (that requires a clear formalization)
- By markup we can encode the text as what it is and not what it seems

For these reasons, procedural markup (WYSIWIG editors) is not suitable > solution: **descriptive markup > XML**

XML (eXtensible Markup Language)

- Is derived from SGML (invented in the 70s by IBM to manage large amounts of textual data without depending on user OS)
- **Is a simplification of SGML (same features, less complexity)**
- Is extensible > not limited to specific elements (as HTML)
- Is determined by W3C Consortium (http://www.w3.org/), from 1998
- Has different uses > can be adapted for the remediation of textual documents
- Can be modified by a text editor (preferably in Unicode)

XML DOCUMENT (2)

An XML document contains:

A **declaration** (this is an XML doc)
Elements and attributes, mixed with:

- ✓ The textual content of the document
- ✓ Editor comments
- ✓ External entities (f.e. digitized images)
- ✓ Namespace elements (such as databases/collections of names)

XML DOCUMENT (3)

- must be **well formed** (on the base of internal XML sintaxis)
- can be validated by a specific **validation schema**

A **schema** contains a definition of encoding elements admitted in the final document, can be created by:

- DTD (Document Type Definition/Declaration)
- schema (W3C schema, RelaxNG)

✓ validation is suitable for complex docs, such as literary texts

✓ can be extended by the extension of the schema

XML DOCUMENT (4)

Each XML doc:

- Begins with a processing instruction line that declares that it is an XML doc
- Contains a finite number of elements, marked by an opening tag and a ending tag `<title>This is a title</title>`
- Attributes can be associated only to the opening tag

  `<title type="main">This is a title</title>`
- Can contain empty elements `<pb/> <gap/>`
- Comments can be delimited by `<!-- comment -->`

XML DOCUMENT example

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE TEI SYSTEM "tei-lite.dtd">

<TEI xmlns="http://www.tei-c.org/ns/1.0">

  <teiHeader>...</teiHeader>

  <text>
    <body>
      <p></p>
    </body>
  </text>

</TEI>
```

TEXT ENCODING INITIATIVE (TEI)

site WWW: http://www.tei-c.org/

“an international and interdisciplinary standard that enables libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation”

Reference: *Guidelines for Electronic Text Encoding and Interchange* (http://www.tei-c.org/Guidelines/)

TEXT ENCODING INITIATIVE (TEI)

Brief history

1987: need of a standard that permits the creation and interchange of textual archives (NY Congress)

- 1990: TEI Guidelines first version (TEI P1)
- 1990-94: funds from NEH, Mellon Foundation, European Community; support of ACH (Association for Computers and the Humanities), ACL (Association for Computational Linguistics), ALLC (Association for Literary and Linguistic Computing)
- 2000: TEI Consortium was born, no profit association for the development of TEI standard
- 2002: TEI P4 version > passage from SGML to XML
- 2007: TEI P5 version, constantly updated

TEXT ENCODING INITIATIVE (TEI)

- better interchange and integration of scholarly data
- support for all texts, in all languages, from all periods
- guidance for the perplexed: what to encode --- hence, a user-driven codification of existing best practice
- assistance for the specialist: how to encode --- hence, a loose framework into which unpredictable extensions can be fitted
- These apparently incompatible goals result in a highly flexible, modular, environment for DTD customization

TEI MODULAR STRUCTURE

Essential modules:

– **tei** define elements' classes, macros and datatypes for all modules

– **header** includes metadata of TEI XML elements

– **textstructure** structural elements for every kind of text

– **core** significant text elements

– more details here (Chapter 2: TEI infrastructure)

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

TEI

Mminimum TEI P5 document is composed by:

- XML declaration and in case *processing instructions*
- TEI Header
- structural elements (book materiality)
- semantic elements (not necessary)

Basic modules: **tei**, **header**, **textstructure**, **core**
(help to model a large amount of different texts)

TEI DOCUMENT examples

```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <?xml-model href="prohd.rng"  schematypens="http://relaxng.org/ns/structure/1.0"  type="application/xml"?>
3   <?xml-model href="prohd.sch" type="application/xml" schematypens="http://purl.oclc.org/dsdl/schematron"?>
4   <TEI xmlns="http://www.tei-c.org/ns/1.0" type="document" xml:id="prohd">
5       <teiHeader>
6           <fileDesc>…
49          </fileDesc>
50          <profileDesc>…
68          </profileDesc>
69          <encodingDesc>…
105         </encodingDesc>
106     </teiHeader>
107     <text>
108         <body>
109             <div>
110                 <pb n="786r" facs="prohd0020_786r.tif"/>
111                 <gap reason="insignificant" quantity="12" unit="lines"/>
112                 <p>
113                     <hi rendition="#b">El Sr. Zayas
114                         <lb/>sobre laventa
115                         <lb/>dela obra del
116                         <lb/>Barón de Hum
117                         <lb/>boldt
118                     </hi>
119                 </p>
120                 <p>Con este motibo el Sr. D.n Andres de Za
121                     <lb/>yas espuso que en las librerias de esta ciudad se estan ven-
122                     <lb break="no"/>diendo en castellano el ensayo politico sobre la Ysla de Cuba
123                     <lb/>escrito por el Baron de Humboldt; que esta obra bajo muchos
124                     <lb/>aspectos apreciabilisimos era sin embargo sobremanera pe
```

TEI DOCUMENT examples